TESIS DOCTORAL

Programa de Doctorado en Psicología

Proactive and reactive control during social information processing in neutral and interpersonal contexts.

Doctoranda

Paloma del Rocío Díaz Gutiérrez

Directora

María Ruz

UNIVERSIDAD DE GRANADA

Departamento de Psicología Experimental

Octubre de 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Paloma del Rocío Díaz Gutiérrez
ISBN: 978-84-1306-410-9
URI: http://hdl.handle.net/10481/58768

Agradecimientos

Just because you've got the emotional range of a teaspoon doesn't mean we all have

H.J. Granger

Quienes me conocen saben que quise ser muchas cosas a lo largo de los años (dependiendo de la serie de turno), que acabé en esto de la Psicología y Neurociencia Cognitiva por casualidad, y que mi mejor habilidad es recordar cosas especialmente irrelevantes. Por eso, el haber llegado hasta aquí no hubiera sido posible sin una serie de personas. Desordenadamente, os doy las gracias (a mi manera) a quienes directa o indirectamente habéis contribuido a que hoy esté escribiendo esta tesis. Dentro tocho.

María, muchísimas gracias por acogerme desde tan pronto y tan bien. Gracias por animarnos siempre a aprender cosas nuevas, aunque al principio sean incomprensibles. Gracias por tu aguante y paciencia ante los experimentos tormentosos. Por las cervezas de los viernes y el salmorejo.

Sam, thank you for letting me be part of your group at the ICN for those three months. I'm extremely grateful for all your help these last years. Annika, thank you too for welcoming me in your office and those afternoons spying from the window.

Gracias a la gente del grupo de Neurociencia Cognitiva, por todas las reuniones sociales y académicas donde he podido aprender un poco de todos (y de todos). Gracias Juan, por acogernos cada diciembre en el fiestódromo. A la gente que ha llenado los pasillos del CIMCYC y ha hecho del trabajo algo más entretenido: Isma, Omar, Itsaso, Carmen…y un largo etcétera. Nuria, por tu breve estancia en el 345 y ocupar voluntariamente el hueco de al lado. Daniel, gracias por esos jueves en el López Correa que mi espíritu competitivo echa tanto de menos. Maika, por la reciprocidad (que existe), por Ed, el Trust Game y compartir estas últimas semanas infernales. Luis (Luis!), gracias por tu insistencia con el futbolín y tus telómeros gordetes, aunque no por hacerme escalar una montaña (o así lo viví) para ir a tu casa a hacer cata de cerveza.

Por la parte técnica gracias también a Juan Carlos por las charlas en el lab y salvar mi ordenador en el momento más crítico. Gracias Jose y Felix, por hacer las 150 horas que he pasado en el escáner (sí, he hecho cálculos) algo más livianas.

También gracias a Fer, por las costishitas y empanadishas, por alterar mi tranquila rutina y ser mi "pibe" favorito. Mil gracias por los ánimos y la ayuda transatlántica. Ma' Enzo, gracias por ser malignamente encantador, por compartir el amor hacia Amy, *you know we're not good*, y por esas historias (¿reales?) tan locas como tú. Espero con ganas nuestro dúo al ukelele y el *Nessun dorma*. Javi, gracias por tu cariño, tus clases de boxeo (mi puño de acero es gracias a ti), por estar siempre al pie del cañón dispuesto a todo y contagiarnos tus ganas de hacer cosas. A los tres, se os hecha mucho de menos.

Por otro lado, gracias a los molledanos de la OJMM, a Borja, Lía, etc, porque gracias a los sábados de ensayos no perdí la cabeza en bachillerato, y seguí con la música. Alonso, quién me iba a decir que una clase de Historia de la Estética Musical (¿?) iba a desencadenar todo esto.

A los Willy Fogs: María, Auro, Dani, Ana y compañía. Gracias por los viajes, los baños nocturnos y los revolcones con la colchoneta en la playa. Alba, gracias por ser mi compañera de conciertos y de cantes (lo dice la Domi no yo) y acompañarme hasta otro país para perseguir nuestros delirios pelirrojos. Cris, gracias por compartir conmigo parte del camino, por la paciencia ante los cambios de agenda y mi obsesión con el mail. Sobretodo, gracias por todos esos viajes en coche donde me dejabas poner la música que quería mientras me dejaba la voz ~~gritando~~ cantando.

Raquel, gracias por estar desde hace tanto y por saberlo todo, aunque llevemos casi 10 años sin apenas vernos. Gracias por el boqueroncito, el "berigüell", los chelistas, tu paciencia con el portón y todo tipo de momentos bochornosos (que no son pocos).

Celia y Belén, gracias por *only 3* y los consejos de sabias. B, gracias por ser mi primera amiga en Granada, conslarme cuando me ~~dejé estafar~~ estafaron y hacerme planear los conjuntos dos meses antes de un evento. Gracias por el pinzaball, pero, sobre todo, por enseñarme la mayor lección que recibiré en esta vida: NO se llevan regalos en bolsas del Mercadona.
C, gracias por ser mi otro limón (porque la media naranja no existe), por aguantarme estos 9 años viviendo juntas, por las incontables horas viendo series y películas (estas sí que no las cuento). Gracias por atreverte a ser la mamá nº2 de Rupert. Gracias por escucharme y conocerme tan bien que no pasa nada cuando te cuento las cosas con años de retraso, porque tú ya lo sabías.

Para ir cerrando (ya queda menos!) tengo que agradecer al 345, a los que estuvieron y están. por ser básicamente mi familia en estos años. Gracias por el amigo invisible y las tartas de cumpleaños compradas a última hora (como debe ser):

Sonia, gracias por seguir educándonos en la distancia, por ser tan bella por dentro y por fuera, por tus *dai* y estornudos japoneses. Carlos, gracias por seguir siendo el "maehtro Carlos" en la distancia, por siempre estar dispuesto a ayudar con todo lo que puedas, y por ser el creador de los *stickers* del 345 (menos ya sabes cual, #noOmbre). A los dos, no sé si me hubiera metido en todo esto si no os hubiera conocido el primer día.

David, gracias por crear esa plantilla de powerpoint de la que todos estamos celosos. Por ser aparentemente perfecto (payaso), esas historias de curas y hostias, y por ser el mejor telonero de Julio Iglesias. Chema, gracias por tu motivación, por tus consejos de scout (aunque no por la amenización musical mientras bajo el monte). Gracias por compartir el amor por el Doctor (*I don't wanna go*).

JE, completo lo que te dije en tu tesis (para darle más *punch*). Gracias por tu paciencia (conmigo y con PAP en general), por estar ahí en cada agobio y escucharme siempre. Gracias de verdad, porque sin tu ayuda y tus scripts no podría haber hecho ni la mitad de lo que hay en esta tesis. Gracias por escuchar todas mis historias incluso cuando no te interesan, y por dejarme contarte el *plot* completo de alguna serie cuando no tengo con quien comentarla. Gracias por los cafopas y Apache. Por todo, menos por ese póster del Madrid.

Alberto, gracias por ser el mejor compañero de mesa que podía pedir, por toda la ayuda y ánimos en las últimas semanas. Gracias por tus conocimientos gatunos, los pedidos de domingo (digo, la comida *real food*) y por acordarte de lo que llevaba ese día (ah no, que era mentira). Gracias por dejarme narrar tus historias embarazosas (reales y ficticias) y por escuchar mis historias nostálgicas, aunque luego te rías de mí. Gracias por aguantarme cuando me comunico con sonidos que nadie entiende, cuando te imito (*oh, lady Ernstwhile*) y cuando proporciono información completamente innecesaria sobre mí. Te doy mi sello.

AP, podría escribir otra tesis narrando nuestras aventuras y desgracias, así que ~~intentaré ser~~ seré breve. Gracias por todas esas horas y emociones variables que hemos compartido a lo largo de 5 años. Gracias por nuestro gran éxito musical *Cimcyctime sadness*, por los hashtags y las tardes programando en la Qarmita. Por encima de todo, gracias por ser mi regresión a la media, el ello a mi superyó, la víctima rubia a la morena; por ir a la par en todo, aunque siempre fueses por delante. Estoy convencida de que no sería como soy ahora si no hubiera compartido estos años de ~~tortura~~ aprendizaje y experiencias contigo (*sounds dramatic but ok*). Gracias por encargarte de llamar a los sitios cuando hace falta porque sabes que ~~me da vergüenza~~ no me gusta. Gracias por PAP. Tra tra.

A mis padres, por los ánimos estos últimos meses y por todo, siempre. Gracias por hacer todo lo posible porque tuviese de todo y lo mejor, por entender (o aguantar, qué remedio) mi reticencia a volver a casa. Porque, siendo a ojos de todo el mundo una mezcla de ambos, sin vosotros literalmente no podría haber llegado hasta aquí y no sería quien soy.

Contents

List of figures and tables

Acronyms

ACC	Anterior Cingulate Cortex
aI	Anterior Insula
ATL	Anterior Temporal lobe
DLPFC	Dorsolateral Prefrontal Cortex
DMN	Default Mode Network
EEG	Electroencephalography
FFA	Face Fusiform Area
FG	Fusiform Gyrus
FIR	Finite Impulse Response
fMRI	functional Magnetic Resonance Imaging
FWE	Family-wise-error
fO	Frontal Operculum
FP	Frontoparietal
HRF	Hemodynamic Response Function
IFG	Inferior Frontal Gyrus
IPL	Inferior Parietal Lobe
IPS	Inferior Parietal Sulcus
ITI	Inter-trial-interval
ITG	Inferior Temporal Gyrus
LG	Lingual Gyrus
LOSO	Leave-One-Subject-Out
lPFC	Lateral Prefrontal Cortex
LSS	Least-Squares-Separate

MCC	Middle Cingulate Cortex
MDN	Multiple Demand Network
MEG	Magnetoencephalography
MFG	Middle Frontal Gyrus
MFN	Medial Frontal Negativity
MM	Mixed Miniblock
mPFC	Medial Prefrontal Cortex
MTG	Middle Temporal Gyrus
MVPA	Multivariate Pattern Analysis
OFC	Orbitofrontal Cortex
PC	Precuneus
PCC	Posterior Cingulate Cortex
PFC	Prefrontal Cortex
PM	Pure Miniblock
ROI	Region of Interest
RSA	Representational Similarity Analysis
SMA	Supplementary Motor Area
SPL	Superior Parietal Lobe
TG	Trust Game
ToM	Theory of Mind
TPJ	Temporoparietal Junction
UG	Ultimatum Game
VLPFC	Ventrolateral Prefrontal Cortex
vmPFC	Ventromedial Prefrontal Cortex

CHAPTER 1

INTRODUCTION

The general aim of the present thesis is to study the neural basis of the control mechanisms that underlie successful navigation in social and non-social scenarios. Humans possess a remarkable set of cognitive skills to behave flexibly according to goals, in adaptation to a complex environment. This repertoire of abilities, known as cognitive control, is crucial to dealing with situations that entrain some difficulty, for instance when we perform novel tasks or where routine procedures lead to divergent responses (Norman & Shallice, 1980). Throughout the years, several theoretical models have characterized this mechanism (Baddeley, 1992; Norman and Shallice, 1980; Posner & Snyder, 1975), emphasizing its contribution in numerous contexts and in presence of inputs of different nature. Thus, a central question in Psychology and Cognitive Neuroscience has been to understand how executive control affords such versatility.

Many studies have examined the implementation of cognitive control through a variety of paradigms, from simple ones like the Stroop task (Stroop, 1935) to complex settings where participants need to follow novel verbal instructions (e.g. González-García, Arco, Palenciano, Ramírez, & Ruz, 2017). Further, the exponential advance in neuroimaging methodologies such as functional Magnetic Resonance Imaging (fMRI) has boosted our understanding of the neural basis of cognitive control. As a result, recent proposals have complemented earlier models (e.g. Desimone and Duncan, 1995; Posner and Petersen, 1990), characterizing in more detail the implication of different brain areas (Braver, 2012; Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008). Moreover, novel analytic approaches like Multivariate Pattern Analysis (MVPA; Haxby, Conolly & Guntupalli, 2014; Haynes, 2015) have led to a refinement of

our experimental questions, providing the tools to examine how the brain represents relevant information to fulfil task demands. This thesis contains three studies employing these methods, and focuses on social and non-social domains of executive control.

The last decade has witnessed a growth in MVPA studies (Woolgar, Jackson, & Duncan, 2016), which show that adjustments in behaviour in accordance to our goals are mediated, at least in part, by changes in the representation of information (Waskom, Kumaran, Gordon, Rissman, & Wagner, 2014). Some studies have examined information coding for current task performance (e.g. (Qiao, Zhang, Chen, & Egner, 2017; Woolgar, Thompson, Bor, & Duncan, 2011) while others have focused on future goals (Gilbert, 2011; Haynes et al., 2007). Yet, complex behaviour requires the ability to act according to a series of plans in a certain order, maintaining both current and future task-relevant information during ongoing task performance. Thus, the first aim of this thesis was to examine, within the same experimental setting, the brain mechanisms underlying the representation of current and future task sets.

Moreover, recent models have emphasized that the brain's ability to implement efficient control relies on a set of regions carrying specific computations. In particular, Dosenbach and colleagues (2008) proposed that control mechanisms consist of two differentiated networks that participate at different temporal scales. On the other hand, Braver (2012) distinguished between two control modes: one that prepares the organism in advance and another that offers late responses to conflict. These models have been tested previously in scenarios of varied complexity (e.g. Dubis, Siegel, Neta, Visscher, & Petersen, 2016;

González-García, Mas-Herrero, de Diego-Balaguer, & Ruz, 2016; Marini, Demeter, Roberts, Chelazzi, & Woldorff, 2016; Palenciano, González-García, Arco, & Ruz, 2019). However, we still do not know if this characterization of control mechanisms can be extended to different domains, specifically, interpersonal scenarios. Previous work on strategic behaviour during social interactions (e.g. Ruz, Madrid, & Tudela, 2013; Ruz & Tudela, 2011) illustrates the importance of understanding control in complex scenarios like interpersonal decisions. Therefore, a key aim of this thesis was to examine whether control mechanisms contribute similarly in social and non-social interactions. To this end, the second goal of this thesis was to study the control networks underlying the sustained maintenance of task sets and their transient response to conflicting information, as well as to examine if these mechanisms varied depending on the nature (social vs. non- social) of our goals.

Likewise, although our interpersonal decisions are accompanied by a sense of rationality, they are biased by evaluative information about others (Díaz-Gutiérrez, Alguacil, & Ruz, 2017). Part of these influences is prompted by the stimuli we perceive: social categories, emotions, or even descriptions about their personality. Importantly, this information biases our choices even when it is not related to people's behaviour (Alguacil, Tudela, & Ruz, 2015; Tortosa, Strizhko, Capizzi, & Ruz, 2013). Thus, when we are faced with different social cues that lead to divergent predictions or when our expectations are not matched, control mechanisms become necessary to detect and implement the appropriate adjustments. In this line, some studies have evidenced the important role of preparatory processes ahead of novel and complex tasks (González-García et al., 2017) and how the expectation of subsequent events modulates brain activation

prior to stimulus onset (Summerfield & De Lange, 2014). Considering the influence of social knowledge in decision-making in interpersonal situations (Díaz-Gutiérrez et al., 2017), the third goal of this thesis was to study how the brain represents personal information about others to expect beneficial or disadvantageous outcomes. Here we examined how these expectations influence social decisions and how our brain responds when predictions mismatch actual events.

In the pages that follow, we contextualize the theoretical framework of the thesis and of the specific research questions we address. We begin with a general description of the different approaches in the study of control mechanisms, and the neural representation of task-relevant information. Then, we move to social interactions, where we highlight the influences that modulate interpersonal decisions. Last, we review the existent work on the contribution of control in these contexts.

I. What is cognitive control? Neural components.

Formally, cognitive control can be understood as the set of mechanisms that guide and regulate thought and behaviour according to goals (Braver, 2012; Miller & Cohen, 2001). From earlier to recent proposals, all of them agree on its flexibility and adaptability, as well as on its influence across multiple domains. In this sense, Braver (2012) distinguished two main control modes in terms of their temporal profiles. This way, control can be implemented in a proactive or reactive manner. First, proactive control anticipates task demands and therefore maintains task-relevant information in an active state. Conversely, reactive

mechanisms are recruited to detect and resolve conflict situations during task performance. In the following section we describe both components and their neural mechanisms.

a. Proactive control

According to Braver (2012), proactive control maintains, in an active state, task-relevant information in accordance to one's internal goals. In this way, this mechanism allows us to prepare for upcoming events. To do this, proactive control adjusts behaviour and stimuli processing according to a task set, which is the configuration of attentional, perceptual and motor processes that are necessary to perform a particular task (Sakai, 2008). The implementation of a task set requires different computations, ranging from the representation of abstract goals, to the specific task rules and the association between stimuli and responses (Rubinstein, Meyer, & Evans, 2001). The study of these processes often employs paradigms where a cue indicates or provides information about the task to perform next (González-García et al., 2016; Monsell, 2003). These paradigms allow examining cue-related activation before the target appears, and also how the brain reconfigures or updates mental sets when the task changes.

The maintenance and representation of task sets are associated with frontoparietal regions that represent task-relevant information at different levels of abstraction (Palenciano, Díaz-Gutiérrez, González-García, & Ruz, 2017). Within these, the prefrontal cortex (PFC) is crucial for proactive control, thanks to its connections with specialized regions that amplify the representation of task-relevant information (Sakai, 2008). Specifically, the

lateral PFC is associated with the maintenance of task rules (Brass & von Cramon, 2004), while more anterior parts (rostral PFC) contain information about abstract goals or intentions (Burgess, Scott, & Frith, 2003; Haynes et al., 2007).

In addition, a set of regions beyond the PFC also contributes to proactive control, including the pre-supplementary motor area (pre-SMA) and parietal cortex. The pre-SMA has been associated with unspecific preparation (Brass & von Cramon, 2004), but also with the suppression of interference from other task sets (Crone, Wendelken, Donohue, & Bunge, 2006; De Baene & Brass, 2014). On the other hand, the superior parietal lobe (SPL) contributes to the activation of task rules (Muhle-Karbe, Andres & Brass, 2014), whereas the inferior parietal lobe (IPL), specially the intraparietal sulcus (IPS), represents stimulus-response associations (Brass & von Cramon, 2004; Muhle-Karbe et al., 2014).

In addition, a cingulo-opercular network conformed by the anterior cingulate cortex (ACC), the anterior insula and frontal operculum (aI/fO), shows sustained activation during an extended time, together with the rostral PFC (Dosenbach et al., 2007; Dosenbach et al., 2006). This observation has been interpreted by ascribing a contribution of these regions in the stable maintenance of task goals (Dosenbach et al., 2008). According to this proposal, the role of the cingulo-opercular would be complementary to the one of a frontoparietal network, which would be in charge of the phasic initiation and adjustment of control.

b. Reactive control

In addition to the anticipation of events, control mechanisms also implement adjustments when needed, usually in reaction to conflict (Braver, 2012). This conflict is usually triggered by events that interfere with information processing, generally caused by the coactivation of incompatible action tendencies (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Reactive control is typically studied with paradigms of interference where irrelevant stimuli dimensions trigger automatic actions that interfere with the response associated with the relevant dimension (Egner, 2008). Among these tasks, the most representative are the Stroop (Stroop, 1935), Flanker (Eriksen & Eriksen, 1974) or Simon (Simon, 1969) tasks. Interference or conflict effects are reflected in declined performance, with slower and less accurate responses for incompatible (or incongruent) compared to compatible (or congruent) events.

In relation to the neural basis of reactive control, one of the most influential proposals has been the "Conflict monitoring hypothesis" by Botvinick et al. (2001). According to this model, reactive control would depend on two processes: conflict monitoring and conflict resolution. Conflict monitoring and detection have been associated with the ACC, especially its dorsal part (dACC; Botvinick, Cohen, & Carter, 2004; Kerns et al., 2004; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004; but see Rushworth et al., 2005). Likewise, the dACC is also sensitive to the amount of conflict (Etkin, Egner, Peraza, Kandel, & Hirsch, 2006). Moreover, the role of this monitoring mechanism would be to signal the presence of conflict to other control areas to make the appropriate adjustments (Botvinick et al., 2001). Here, the PFC is associated with a key role

in conflict resolution. This brain region would bias the competition for representation in sensory areas between relevant and irrelevant stimuli (Desimone & Duncan, 1995), to enhance the processing of task-relevant information (Miller & Cohen, 2001). In this way, Egner & Hirsch (2005) showed that the dorsolateral PFC (DLPFC) facilitated the processing of relevant stimuli in sensory regions during conflict. On the other hand, the ventral portion of lateral PFC (VLPFC) has been associated, among other roles, with the inhibition of irrelevant information, in concert with the aI and pre-SMA (Levy & Wagner, 2011). Additionally, parietal regions may also play a role in the inhibition of distractors during conflictive situations (Marini et al., 2016). As we can see, frontoparietal regions are also important for reactive adjustments of control. This highlights the flexibility of these regions to implement control in a proactive or reactive manner (Marini et al., ,2016), depending on task needs.

c. Control networks and the representation of task goals.

Altogether, the extensive work in control mechanisms agrees on the important role of frontoparietal regions. These are also known as the Multiple Demand network (MD; Duncan, 2010) and they are related to demanding and effortful scenarios. On top of this, these areas code task-relevant information during performance (Qiao et al., 2017; Waskom et al., 2014; Woolgar et al., 2011b) and this representation shows adaptability to context demands (adaptive coding; Woolgar et al., 2016). Importantly, the activation increase in these regions has been typically associated with the deactivation of the so-called Default Mode Network (DMN; Fox et al., 2005; Raichle, 2015). At first this was interpreted as an indicator of the absence of a functional role in experimental tasks, at least in

those that entrain some difficulty. However, recent work indicates that the DMN network also codes task sets (Crittenden, Mitchell, & Duncan, 2015). Further, several studies in prospective memory have related the medial PFC (mPFC), one of the main nodes of the DMN, to the representation of future goals (Gilbert, 2011; Momennejad & Haynes, 2013; Momennejad & Haynes, 2012). The mixed evidence raises a question about whether frontoparietal regions and the DMN have complementary roles in control, underlying the representation of ongoing vs. future task sets, respectively.

In this section we have described how control mechanisms prepare us and provide adjustments during task performance. In addition, we have shown that task-relevant information coding is crucial to the flexibility of control. However, these studies are carried out within experimental settings that do not consider the intricate aspects of our daily social lives. Therefore, in the next section we focus on a special type of social behaviour in which we are frequently involved: interpersonal decisions.

II. Social decision-making

Social interaction is one of the most outstanding components of human behaviour. Humphreys (1976) stressed the relevance of social contexts, manifesting that the complexity of our collective environment is what promoted our sophisticated cognition. Our rich social life involves continuous interactions with other people. A large part of these involves making decisions about them, where we need to consider information to predict their likely behaviour (Díaz-Gutiérrez et al., 2017). Thus, apart from considering the value of different

options and potential rewards, social decisions call for predictions about others' mental states (Lee & Harris, 2013). Also, social decisions entail interactions between strategic processing and emotion, since both can be insightful (Lee & Harris, 2013). In some cases, emotion can prevent us from damaging others (Greene & Haidt, 2002), whereas overriding social biases may be the appropriate response in some situations (Rilling & Sanfey, 2011).

Although there are different types of social decisions (Lee & Harris, 2013), we focus here on strategic interaction, where previous studies have evidenced the critical role of control mechanisms. In this type of decisions, we depend on our choices as well as others' (Rilling & Sanfey, 2011; Sanfey, 2007). Generally, strategic interactions are examined through economic games. Since part of the focus of this thesis relies on the study of social decision in strategic scenarios, we describe the two game paradigms that we employ in this thesis.

a. Paradigms to study social decision-making

Economic games are based on Game Theory, which comprises a variety of models that try to explain decision-making in an interpersonal context (Rilling & Sanfey, 2011). These games are useful tools to examine social prediction in interpersonal choices, since they are easy and can be adapted to neuroimaging procedures. One of the most popular is the Trust Game (TG; Camerer, 2003), which is used to study reciprocation behaviour. Here, a participant acts as an investor deciding whether or not to share a certain amount of money with a partner (trustee). If the investor decides to cooperate and share the money, that amount is multiplied and transferred to the trustee. Then, the trustee decides to

reciprocate or not with the investor. In the first case, both earn half of the total money, and in the second scenario the trustee gets everything, whereas the investor loses the initial sum. Thus, in this context cooperation between both parts is the best strategy. Another popular paradigm is the Ultimatum Game (UG; Güth, Schmittberger, & Schwarze, 1982), which explores people's response to fairness. In this case, one participant acts as a proposer, who decides how to split a sum of money. Then the other player acts as a responder, choosing to accept or reject the offer. In the first scenario, both parts earn their corresponding split, whereas in the opposite case none of them gains anything. Thus, here the most rational strategy would be to always accept the offers.

In this way, these paradigms allow the study of different phenomena that influence interpersonal decisions, from reciprocation behaviour to response to fairness. In both cases it is key to predict what people will do and act accordingly to make the most efficient decision. Prediction of people's intentions and traits is essential for making decisions about them, since it helps us to anticipate their most likely behaviour. Therefore, it is important to understand how this social information is represented in the brain and to examine how it is coded during interpersonal decisions. For this reason, we now describe the main regions associated with the representation of social knowledge and how this information can be obtained.

b. Acquisition of social knowledge and its representation

As experienced as we are navigating social life, we have sharpened skills to infer or acquire several types of knowledge about others that we can use to predict

their most likely behaviour. Different mechanisms make it possible to obtain such social information. One of these is social categorization. Although a powerful strategy to facilitate perception, categorization is related to stereotypes that generate specific expectations about how people might be, which may lead to prejudice (Freeman & Ambady, 2011). Moreover, emotions are a form of social communication and motivate social behaviour (Adolphs & Anderson, 2013). People make judgements about facial expressions, associating positive emotions with trustworthiness and negative ones with unreliability (Oosterhof & Todorov, 2008). Further, social knowledge can be learned through direct experience when we interact with others (Koster-Hale & Saxe, 2013), but we can also predict people's behaviour and generate expectations based on personal descriptions about others (Tamir & Thornton, 2018).

A large part of these judgements relies on people's faces, which contain a large amount of information about them (e.g.: age, sex, identity, emotion, race). Haxby, Hoffman & Gobbini (2002) proposed a model for face perception. This takes place in a core system in charge of the visual analysis of faces that encompass the inferior occipital gyrus extending to the fusiform gyrus (also known as face fusiform area, FFA; Kanwisher, McDermott, & Chun, 1997) and superior temporal sulcus. The model also includes a set of extended areas associated with different processes: the intraparietal sulcus and frontal eye fields for processing gaze direction or head position, the anterior temporal lobe for retrieving semantic knowledge (e.g. names) and the amygdala and insula for emotional expressions.

Importantly, recent work has shown that face-processing areas also contain higher-level information such as social categories (Stolier & Freeman, 2016). These authors showed the intersection of social categories in fusiform gyrus (FG) as well as in a high-level region such as the orbitofrontal cortex (OFC), which highlights the importance of dynamics between top-down and bottom-up mechanisms in person perception (Stolier, Hehman, & Freeman, 2018). On the other hand, Satpute & Lindquist (2019) gathered data from different studies and showed that the DMN underlies the representation of discrete experience of emotions. Additionally, these regions have a role in social priming and social learning (Meyer, 2019). Also, the connectivity within DMN regions facilitates the encoding of social information, which relates to subsequent thinking about others and facilitates the encoding of social information (Meyer, Davachi, Ochsner, & Lieberman, 2018; Meyer & Lieberman, 2016).

Overall, several studies agree on the existence of a "social or mentalizing brain" that represents knowledge about other people, their mental states or other type of social information (Frith, 2007). Among the regions of this "social brain" we can find the mPFC, temporoparietal junction (TPJ), precuneus (PC), posterior cingulate cortex (PCC) or the anterior temporal lobe (Meyer, 2019; Saxe, 2006). Importantly, some studies (e.g. Ma et al., 2012) have observed the implication of both mentalizing and control-related regions when people's behaviour was inconsistent with their assigned traits.

After describing the representation of social knowledge and its relevance in the prediction of people's behaviour, we move to examine how social information influences social decisions.

c. Influence of social information in interpersonal decisions

As we introduced earlier, several sources of information modulate interpersonal decisions. Among them, emotional expressions are related to trustworthiness judgements, which influence social behaviour. This way, previous studies have shown that positive expressions elicit cooperation and trust, while negative ones are associated with uncooperative decisions (Oosterhof & Todorov, 2008). In addition, emotions have an automatic effect and elicit "emotional conflict" (Díaz-Gutiérrez et al., 2017). That is, where people need more time to choose a course of action contrary to the one elicited by irrelevant and non-predictive emotions (Alguacil et al., 2015). This is also the case when emotions do not predict their natural consequences (happiness = positive outcome, anger = negative output), where again people need more time to select an option (Ruz & Tudela, 2011).

In other scenarios, we do not have any physical input from which to extract social knowledge. However, we can still obtain personal information about others that guide our decisions, even if they are unrelated to people's actual behaviour. First, closeness with others is a factor that impacts our decisions, where people are more cooperative with close friends than with unfamiliar partners (Fareri, Chang, & Delgado, 2015). Conversely, there are situations where we do not have any information about others. Here, however, people can learn their partners' traits during interaction (Telga, de Lemus, Cañadas, Rodríguez-Bailón, & Lupiáñez, 2018; Tortosa et al., 2013), or in advance (e.g. Gaertig, Moser, Alguacil, & Ruz, 2012). In these scenarios a series of studies have shown that people reject more offers coming from morally flawed partners

than those described positively (Delgado, Frank, & Phelps, 2005; Gaertig et al., 2012; Moser, Gaertig, & Ruz, 2014; Ruz, Moser, & Webster, 2011).

Overall, these influences involve a set of mechanisms that guides our behaviour towards decisions that are not always the most successful. In these scenarios it is necessary the regulation of behaviour towards the most efficient choice. Thus, we now examine the work so far on control mechanisms during social decisions.

d. Control mechanisms in social decisions

We have just described some examples of the sources of social knowledge that can modulate decisions. Further, these influences take place even when this information is not predictive of people's behaviour. Moreover, there are scenarios where we encounter more than one type of information, which may lead to divergent predictions. In these circumstances the regulation of behaviour towards the optimal choice is required to focus on the relevant information and ignore the irrelevant one.

The conflict between different sources of information during social decision-making has been associated with mechanisms similar to reactive control in non-social settings. For instance, both the ACC and lateral PFC are recruited to suppress automatic responses related to irrelevant emotional expressions (Ruz & Tudela, 2011). Importantly, the ACC is coupled with further control regions during conflict between social dimensions, manifesting its pivotal role triggering conflict resolution (Alguacil et al., *unpublished*; Ruz & Tudela, 2011). Moreover, the predictions about others' traits or mental states generate expectations about

their behaviour, which can be congruent or not with such predictions. For instance, in an interpersonal scenario we can expect cooperative and fair behaviour. However, sometimes this is not the case, which recruits again control mechanisms. In this line, Sanfey and colleagues (2003) associated DLPFC activation with the regulation of social behaviour, while the ACC signalled conflict when participants encountered disadvantageous outcomes. Conversely, when expectations conflict with observations, the VLPFC has been associated with the alignment of choices with previous predictions (Fouragnan et al., 2013), which could suggest the retrieval from memory of the relevant information according to goals (Bunge, 2004). Overall, regulation processes are needed to adjust our behaviour during interpersonal interactions. However, it is still not clear yet the mechanisms that underlie the maintenance of social information during decisions and how these task sets deal with potential conflict between different sources. Another remaining issue is how the brain represents specific predictions about others to guide decisions. On the work contained in this thesis, we intend to provide responses to these issues, aiming to a better understanding of the role of control mechanisms in the representation of social information and the implementation of adjustments triggered by conflict, in a variety of neutral and interpersonal contexts.

CHAPTER 2

GOALS AND HYPOTHESES

In the previous introductory chapter, we have described how different control mechanisms flexibly guide our behaviour and how these are represented in our brain. Whereas there are numerous studies on these phenomena, most of them focus on simple stimuli or paradigms (but see Egner et al., 2008; González-García et al., 2017; Palenciano et al., 2019b). However, in daily life, we are constantly exposed to stimuli that have a social component. This makes the understanding of how control mechanisms operate when they work with social information crucial, both in neutral contexts and in interpersonal scenarios, where they have an important predictive function. Neuroimaging methods such as fMRI allow examining the neural basis of these phenomena. In addition, analytic methodologies such as MVPA (Haxby, Conolly & Guntupalli, 2014; Haynes, 2015) complement traditional univariate approaches for a better comprehension of the neural representation underlying information coding at a fine-grained level. Therefore, the main aim of this thesis is to study the brain mechanisms underlying the representation of social task-relevant information and interference resolution, both in neutral and in interpersonal contexts.

In particular, we carried out three fMRI experiments to answer specific questions about control mechanisms:

1) Is there a differential role of frontoparietal regions and the DMN in the representation of task-relevant information that depends on the moment when such information is needed? Can this be extended to social stimuli?
2) Do control networks act at different timescales during a social decision task? How is relevant information actively maintained and how different social and non-social dimensions interact with each other?

3) How do we generate expectations about others and in which way this biases our behaviour? Can we decode the specific content of such expectations? What happens when these predictions are not matched by evidence?

2.1. Representation of social information in non-social contexts. Experiment I.

In the introduction, we have discussed the importance of understanding control mechanisms, where a large variety of studies show a set of regions underlying the representation of task-relevant information. This role has been extensively associated with a frontoparietal (or MD) network, which is engaged during effortful or difficult tasks. However, recent studies have also stressed the relevance of regions related to the DMN in the representation of relevant information, especially the representation of delayed task intentions. Therefore, it is not clear yet whether regions of the DMN, especially its main node, the mPFC, have a role in adjusting our behaviour together with cognitive control regions and if different networks underlie task-relevant information depending on when it is needed.

To answer these questions, we ran an fMRI study with a task-switching paradigm where participants had to make sequential categorization judgements (emotion, gender, race) on faces. We decided to employ faces as stimuli, given their social significance and their representation in specialized regions. The experiment was divided into miniblocks, most of them with two sequential judgement tasks. In these scenarios, during the execution of the first task,

participants needed to maintain the relevant information for both the initial and intended tasks. Therefore, with this study, we were able to assess how task demands influence performance, and to what extent the role of regions previously associated with the representation of task-relevant information can be extended to social stimuli. More importantly, we examined how different task contexts impact the representation of faces, that is, whether current vs. delayed information affect their representational format. In line with previous work, we hypothesized that current task sets could be decoded from MD regions and intended ones from areas associated with the DMN, mainly the mPFC.

2.2. Influence of social information during interpersonal decisions.

2.2.1. Control and interference with social and non-social information during a Trust Game. Experiment II.

In the introduction, we described how control can act at different timescales to sustain task-relevant information and transient adjustments of behaviour. However, it is not clear whether these mechanisms are involved similarly during social scenarios, and if they differ depending on the nature (social vs. non-social) of the relevant information.

To address these issues, we employed a Trust Game, where people needed to choose to share or not an amount of money with their partners, a context where cooperation and trust are crucial to make the most efficient decision. In this scenario, we studied sustained and transient activations of previously described control networks during conflictive situations and their differential involvement depending on the nature (social vs. non-social) of the relevant

information that predicted the outcome of the interaction. This allowed us to examine how these control networks represent two types of interpersonal cues (vs. a non-social one) and their interference. In line with the dual model by Dosenbach and colleagues (2008), we expected to observe sustained activation of cingulo-opercular regions associated with task blocks and frontoparietal areas related to phasic control of interference between relevant and irrelevant dimensions.

2.2.2. Neural representation of social expectations during interpersonal decisions. Experiment III.

The impact that social knowledge has on social interactions has been highlighted by several studies. For instance, personal information helps us predict what others are going to do, which affects choices even when it is not related to their actual behaviour. The way these expectations are represented in the brain and what happens when these predictions are violated it is not known.

We carried out a third study to examine (1) the neural representation of valenced personal information that generates social expectations, (2) how these modulate future decisions and (3) the impact of mismatches between predictions and partners' behaviour. We employed fMRI combining univariate and MVPA approaches during a modified Ultimatum Game, where participants acted as responders and had to decide whether to accept or reject fair and unfair offers made by different partners, after receiving information about their moral traits. Here, we predicted that expectations about the partners in the game would be represented in brain areas associated with social cognition and priors

in decision-making, and that we would observe different patterns depending on the valence of such expectations (Lindquist et al., 2015). In addition, the violation of participants' expectations would be related to activation in control-related regions.

The experimental series have been structured in different chapters. Chapters 3 to 5 verse on the Experiments I, II and III, respectively. We examined the representation of social information in a neutral context in Experiment I, and its influence on interpersonal decisions in Experiments II and III. These chapters are each organized with the sections Abstract, Introduction, Methods, Results and Discussion (all references are combined and included at the end of the document). Last, we conclude with a General Discussion section with a summary and implications of the results, in relation to the field of cognitive control.

CHAPTER 3

EXPERIMENT 1

The content of this chapter is published as Díaz-Gutiérrez, P.; Gilbert, S.; Arco, J.E.; Sobrado, A. & Ruz, M. (2020). Neural representation of current and intended task sets during sequential judgements on human faces. *Neuroimage, 204*, 116219. doi: 10.1016/j.neuroimage.2019.116219.

3.1. Abstract

Engaging in a demanding activity while holding in mind another task to be performed in the near future requires the maintenance of information about both the currently-active task set and the intended one. However, little is known about how the human brain implements such action plans. While some previous studies have examined the neural representation of current task sets and others have investigated delayed intentions, to date none has examined the representation of current and intended task sets within a single experimental paradigm. In this fMRI study, we examined the neural representation of current and intended task sets, employing sequential classification tasks on human faces. Multivariate decoding analyses showed that current task sets were represented in the orbitofrontal cortex (OFC) and fusiform gyrus (FG), while intended tasks could be decoded from lateral prefrontal cortex (lPFC). Importantly, a ventromedial region in PFC/OFC contained information about both current and delayed tasks, although cross-classification between the two types of information was not possible. These results help delineate the neural representations of current and intended task sets, and highlight the importance of ventromedial PFC/OFC for maintaining task-relevant information regardless of when it is needed.

3.2. Introduction

The selection and maintenance of relevant information is critical for our ability to pursue complex and hierarchically organized goals. In cases where we hold delayed intentions that need to be fulfilled later on (also known as prospective memory; Kliegel, McDaniel, & Einstein, 2008) or when we perform sequential tasks, it is necessary to represent the currently-active task and, in addition, the one to be performed later on. It is also important to switch flexibly from one task set to another (e.g. Monsell, 2003). Some studies have examined the neural representation of currently-active task sets in frontoparietal areas (e.g. Waskom, Kumaran, Gordon, Rissman, & Wagner, 2014; Woolgar, Thompson, Bor, & Duncan, 2011), while others have investigated the representations of delayed intentions suggesting a key role of medial prefrontal cortex (mPFC) in combination with more posterior areas (e.g. Gilbert, 2011; Haynes et al., 2007; Momennejad & Haynes, 2013). However, no previous study has examined the representation of current and intended task sets within a single experimental paradigm. This combination allows to investigate the extent to which currently-active and intended future task sets are represented in overlapping versus distinct brain networks, and also to contrast their activation patterns directly. Furthermore, previous studies have focused on representations of rather simple stimuli (i.e. geometric figures, objects, words, etc.; Crittenden, Mitchell, & Duncan, 2015; Waskom et al., 2014; Woolgar et al., 2011b), so it is not clear how well these findings generalize to more complex stimuli, such as human faces. In this study, we employed social categorization dual-sequential judgments on human faces to investigate the common and differential representation of current and delayed tasks.

The influence of maintaining an intended task-set on current task performance has previously been investigated with behavioural methods. These studies highlight how performance declines with an increment the number of tasks that need to be maintained, showing that the representation of two tasks simultaneously is more demanding compared to one task only. For instance, Smith (2003) found that participants performed an ongoing task more slowly when they held in mind a pending intention, compared with performing the ongoing task alone. This behavioural effect is accompanied by changes in pupil dilation (Moyes, Sari-Sarraf, & Gilbert, 2019), which also serves as an indicator of task demands (see van der Wel & van Steenbergen, 2018). Further, dual-task costs have also been manifested in task switching paradigms, where participants must switch between two active task-sets (Monsell, 2003; Rogers & Monsell, 1995). Even when the same task is repeated from the previous trial, responses are slower and less accurate during mixed blocks (where more than one task is relevant) compared with pure blocks consisting of just one task (Marí-Beffa, Cooper, & Houghton, 2012).

Results at the neural level also indicate that the maintenance of two tasks compared with one alters activity in specific brain regions. Several studies have shown that a set of "task-positive" regions increase their activation during demanding tasks (also known as the Multiple Demand network, MD; Duncan, 2010). This network is also sensitive to cognitive load, increasing its sustained activation as task complexity is raised (Dumontheil, Thompson, & Duncan, 2011; Palenciano, González-García, Arco, & Ruz, 2019; but see Tschentscher, Mitchell, & Duncan, 2017). Among these areas, the lateral prefrontal (lPFC) and parietal cortices play a prominent role during dual-task performance. Both increase their activation during task-switching trials while anterior

PFC shows sustained activation during task-switching blocks (Braver, Reynolds, & Donaldson, 2003). Similarly, others (Szameitat, Schubert, Müller, & Von Yves Cramon, 2002) have shown involvement of lPFC during dual-task blocks, proportionally to task difficulty, during simultaneous and interfering task processing. Further, some studies have employed multivoxel pattern analysis (MVPA) to show how these frontoparietal (FP) regions code current task sets (Palenciano, González-García, Arco, Pessoa & Ruz, 2019; Qiao, Zhang, Chen, & Egner, 2017; Waskom et al., 2014; Woolgar, Hampshire, Thompson, & Duncan, 2011) and how the representation of task-relevant information in these areas increases with task demands (Woolgar et al., 2011a).

Traditionally, the role of FP regions has been opposed to "task-negative" areas, initially linked to decreased activity during effortful task performance (Fox et al., 2005), although recent studies suggest that it has a much broader role. This Default Mode Network (DMN) includes the ventro/dorsomedial PFC, orbitofrontal cortex (OFC), precuneus/posterior cingulate, inferior parietal lobe (IPL), lateral temporal cortex, and hippocampal formation (Buckner, Andrews-Hanna, & Schacter, 2008; Raichle, 2015). However, recent studies have qualified this view, showing that these regions also represent task-relevant information in different contexts (e.g. Crittenden et al., 2015; González-García et al., 2017; Palenciano et al., 2019a; Smith, Mitchell, & Duncan, 2018). Moreover, functional connectivity approaches have shown that the strength of connectivity among task-negative regions during a working memory task is associated with better performance (Hampson, Driesen, Skudlarski, Gore, & Constable, 2006). Similarly, Elton and Gao (2015) observed that the dynamics of connectivity among DMN regions during task performance were also related to behavioural efficiency. Altogether, the literature suggests a clear involvement of FP areas in the representation

of current task-related information and highly demanding tasks. Conversely, the role of the DMN is less clear. Although it shows decreased activation during demanding tasks, its dynamics are also related to behaviour, and contain task information in different contexts. This suggests that these regions play a role in the representation of task-relevant knowledge.

Further, one of the main nodes of the DMN, the medial prefrontal cortex (mPFC) has an important role in the representation of intended behaviour during both task-free situations (Haynes et al., 2007) and delays concurrent with an ongoing task (Gilbert, 2011; Momennejad & Haynes, 2012; Momennejad & Haynes, 2013). This area also plays a role when holding decisions before they reach consciousness (Soon, Brass, Heinze, & Haynes, 2008). The evidence from studies of delayed intentions has led to suggested dissociations between the role of lateral and medial PFC (associated with the task-positive and task-negative networks, respectively). Momennejad & Haynes (2013) directly compared the representation of future intentions during delays with and without an ongoing task, and found that while the lPFC had a general role of encoding intentions regardless of whether there was or not an ongoing task during the delay, the mPFC was involved when the delay period was occupied by an ongoing task. Alternatively, Gilbert (2011) could not find encoding of delayed intentions in the lPFC but they did in the mPFC, suggesting that the former may play a content-free role in remembering delayed intentions while the latter would represent their specific content. However, these studies vary in the abstraction of the task rules employed. While Gilbert (2011) aimed to decode specific visual cues and responses, others focused on the anticipation of abstract task sets, such as arithmetic operations (addition vs. substraction; Haynes et al., 2007), or parity vs. magnitude judgements (Momennejad &

Haynes, 2012; Momennejad & Haynes, 2013). This difference in abstraction could impact the brain region (lPFC vs. mPFC) maintaining information about future intentions (Momennejad & Haynes, 2013). Further, these studies also vary in whether the retrieval of the intended task was cued (Gilbert, 2011) or self-initiated (Momennejad & Haynes, 2012; Momennejad & Haynes, 2013). Therefore, although these studies have studied the representation of intentions in a variety of experimental settings, they have not directly addressed how the representation of a future task set may differ from the representation of an ongoing task that is currently being performed.

In addition, the studies so far have employed mainly non-social stimuli. In this context it is worth noting that the DMN has also been related to processes relevant in the social domain (Buckner & Carroll, 2007; Mars et al., 2012; Spreng, Mar, & Kim, 2008). For instance, engagement of the DMN during rest is related to better memory for social information (Meyer, Davachi, Ochsner, & Lieberman, 2018). Facial stimuli are an important source of social knowledge, which is represented in a set of regions including the fusiform gyri (FG; Haxby, Hoffman, & Gobbini, 2000; Kanwisher & Yovel, 2006). This FG also shows different neural patterns distinguishing social categories (Kaul, Ratner, & Van Bavel, 2014; Stolier & Freeman, 2017). Similarly, the representational structure of social categories is altered by personal stereotypes both in the FG and in higher-level areas such as the OFC (Stolier & Freeman, 2016), which is also linked to the representation of social categories such as gender, race, or social status (Gilbert, Swencionis, & Amodio, 2012; Kaul, Rees, & Ishai, 2011; Koski, Collins, Olson, & Hospital, 2017) and the integration of contextual knowledge during face categorization (Freeman et al., 2015). Likewise, during predictive face perception, the FG coactivates with and receives top-down influences from dorsal and ventral mPFC (e.g. Summerfield

et al., 2006), which in turn have also been implicated on judgements about faces (Mitchell, Macrae, & Banaji, 2006; Singer, Kiebel, Winston, Dolan, & Frith, 2004). Therefore, given the special properties and influence of social information gathered from faces, understanding how task-relevant current and delayed information may be represented when it pertains to social information is important to extend and complement previous findings.

In the current fMRI study, we employed a dual-sequential categorization task, where participants had to discriminate between features of three dimensions of facial stimuli and had to maintain for a period of time both the initial ongoing task and an intended one. In particular, we studied how demands (one vs. two sequential tasks) influence performance, and hypothesized that high demand would be associated with worse performance alongside with activation in frontoparietal regions, especially the lPFC. To examine the brain regions containing fine-grained information about both current and intended tasks we employed MVPA. Unlike traditional univariate methods, where the mean activation in a set of voxels is compared between conditions, MVPA focuses on the spatial distribution of activations. Here, a classifier is trained to distinguish response patterns associated with different experimental conditions (i.e. stimuli categories, cognitive states, etc.) in a certain brain region. If the trained classifier is able to predict the patterns of independent data, there is indication that the brain area under study represents specific information about those conditions. Thus, MVPA allows to examine finer-grained differences in how information is represented in the brain (for reviews see Haxby, Connolly, & Guntupalli, 2014; Haynes, 2015). In this work, we aimed to study how an intended task set might be represented differently from a currently-active ongoing task. For that reason, we focused on the initial pre-switch period, when the

current task is being performed before switching to the intended task. Specifically, we performed separate analyses to decode: 1) the task currently being performed, regardless of the intended future task; 2) the task intended for the future, regardless of the current task (henceforth: "initial task" and "intended task", respectively). Given the extensive literature associating FP areas to the representation of task-relevant information (Qiao et al., 2017; Waskom et al., 2014; Woolgar et al., 2011a, 2011b), we expected to decode the initial relevant task in MD regions and the intended one in "task-negative" regions, especially the mPFC, in line with previous studies showing its role in prospective memory (Gilbert, 2011; Haynes et al., 2007; Momennejad & Haynes, 2012; Momennejad & Haynes, 2013) .

3.3. Methods

3.3.1. Participants

Thirty-two volunteers were recruited through adverts addressed to undergraduates and postgraduate students of the University of Granada (range: 18-28, M = 22.5, SD = 2.84, 12 men). All of them were Caucasian, right-handed with normal or corrected-to-normal vision and received economic remuneration (20-25 Euros, according to performance) in exchange for their participation. Participants signed a consent form approved by the Ethics Committee for Human Research of the University of Granada.

3.3.2. Apparatus and stimuli

We employed 24 face photographs (12 identities, 6 females, 6 black; 3 different identities per sex and race) displaying happy or angry emotional expressions, extracted from the NimStim dataset (Tottenham et al., 2009). E-Prime 2.0 software (Schneider,

Eschman, & Zuccolotto, 2002) was used to control and present the stimuli on a screen reflected on a coil-mounted mirror inside the scanner.

3.3.3. Design and procedure

Participants had to perform a series of categorization tasks where they judged either the emotion (happy vs. angry), the gender (female vs. male) or the race (black vs. white) of series of facial displays. These tasks were arranged in miniblocks, which could each contain one (Pure Miniblock; PM) or two sequential categorization tasks (Mixed Miniblock; MM). At the beginning of each miniblock, participants received instructions indicating the number of tasks to perform (1 vs. 2) and their order and nature (Emotion, Gender and/or Race), as well as the key-response mappings. Thus, for PMs, the initial instruction indicated the one task that had to be performed during the whole miniblock. Conversely, for MMs, the instruction indicated two tasks, where the participant had to change from the first to the second task at a certain point of the miniblock. After the instruction, a coloured (blue or red) fixation point appeared on the screen, followed by a facial display (see Figure 1). Participants were told that, during MMs, they had to switch tasks when the fixation changed its colour (from blue to red or vice versa). Once it switched, they had to continue doing the second task until the end of the miniblock. To equate the perceptual conditions across blocks, the fixation colour also changed during PMs, although participants were told to ignore this change.

Hence, in each MM there was an initial task (first task to perform), an intended task (second task to perform) and an ignored task (non-relevant for that miniblock). Importantly, our main fMRI analyses focused on the period before the switch, while participants needed to represent both the initial task and the intended one. Task switches

were evenly spaced across the miniblock, from trial 1 to 12. This allowed us to decorrelate brain activity associated with the pre-switch period, post-switch period, and the switch itself. In total, there were 9 different types of miniblocks: 3 pure (emotion [EE], gender [GG], race [RR]) and 6 mixed (emotion-gender [EG], emotion-race [ER], gender-emotion [GE], gender-race [GR], race-emotion [RE], race-gender [RG], see Figure 2). Across the experiment, pure miniblocks appeared 8 times each, while every type of mixed miniblock was repeated 12 times. The presentation order of the miniblocks and the assignment of response options (left or right index) were counterbalanced within each run. Additionally, to avoid response confounds in the analyses, response mappings changed between runs. Thus, for each participant odd and even runs had the opposite response mappings.

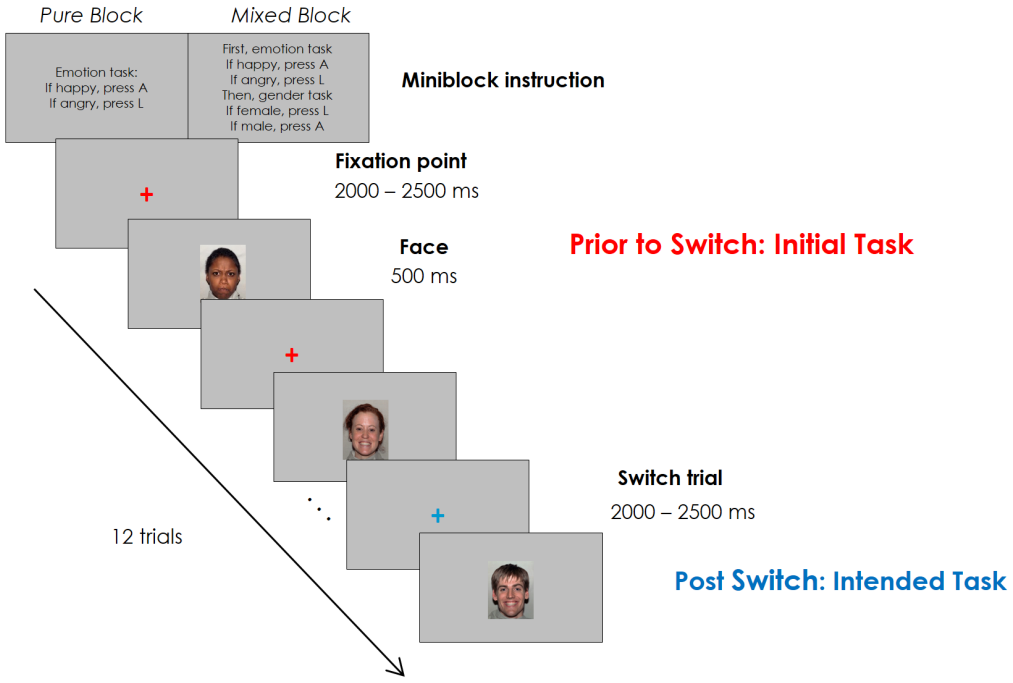

Figure 1. Display of the paradigm. Example of a miniblock and sequence of trials. Inter-trial-interval (ITI) duration = 2-2.5 s.

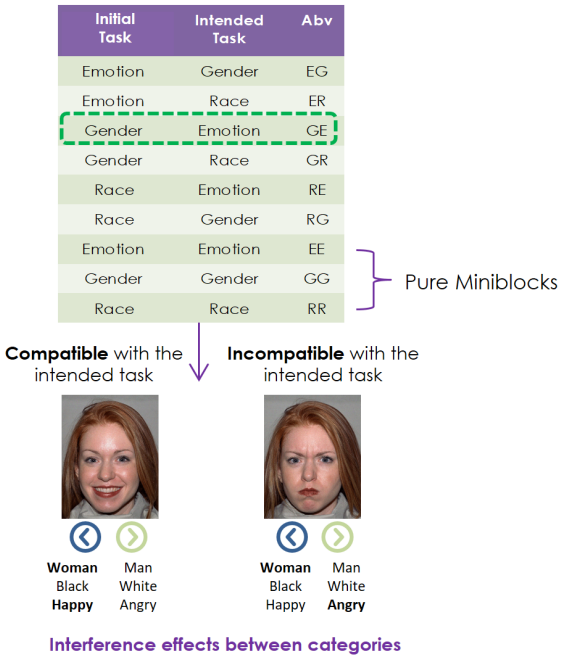

Figure 2. Top: All possible combinations of miniblocks, depending on the initial and intended tasks, and their abbreviation (Abv). Bottom: Example of interference between initial and intended categories in a Gender-Emotion (GE) miniblock.

Participants performed a practice block to learn the different tasks and the response mappings. They were required to obtain a minimum of 80% accuracy at this practice block prior to entering the scanner. The sequence of each miniblock was as follows: First, the instruction slide presented the task/s to perform (Pure: 1, Mixed: 2), and the response mappings (right/left index), during 5 s. Then, a sequence of 12 trials appeared. In each of them, a fixation point (blue or red, counterbalanced) lasting 2-2.5 s (inter-trial-interval; ITI; in units of 0.25 s, randomly assigned to each fixation) was followed by a facial display of 0.5 s. The fixation for the switch trial lasted on average 2.24 s (SD = 0.022; all participants within a range of ± 2.5 standard deviations). The experiment consisted of 1152 trials, arranged in 96 miniblocks (72 mixed and 24 pure), distributed in 12 scanning runs. Hence, each run consisted of 8 miniblocks (6 mixed and 2 pure). Each type of miniblock was repeated 12 times, once per run, and each time the switch

occurred on a different trial. Presentation order and switch point were counterbalanced through the experiment, to ensure that each switch occurred on every possible trial for each type of miniblock and that each identity was associated the same number of times with the switch. In total, the fMRI task lasted for 60.8 min.

In addition, we also studied the interference between tasks. Since in MMs the participant had to perform two tasks sequentially, the established stimulus-response association could be compatible or incompatible between the current and intended task, depending on the specific target face. For instance, the gender task could have a stimulus-response (S-R) association of female-left/male-right and the S-R in the emotion task might be happy-left/angry-right. Therefore, in a GE miniblock, during the gender task, participants could encounter happy female faces (both the initial gender and intended emotion tasks would require the same response: left) or angry female faces (the initial gender task would lead to response with the left index and the emotion with the right one). Thus, the former would be an example of compatibility between initial and intended tasks, whereas the latter would entail incompatibility (see Figure 2).

3.3.4. Image acquisition and preprocessing

Volunteers were scanned with a 3T Siemens Magnetom Trio, located at the Mind, Brain and Behavior Research Center (CIMCYC) in Granada, Spain. Functional images were obtained with a T2*-weighted echo planar imaging (EPI) sequence, with a TR of 2.210 s. Forty descending slices with a thickness of 2.3 mm (20% gap) were extracted (TE = 23 ms, flip angle = 70 º, voxel size of 3x3x2.3 mm). The sequence was divided into 12 runs, consisting of 152 volumes each. Afterwards, an anatomical image for each participant was acquired using a T1-weighted sequence (TR = 2500 ms; TE = 3.69 ms;

flip angle = 7º, voxel size of 1 mm^3). MRI images were preprocessed and analysed with SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12). The first 3 images of each run were discarded to allow the stabilization of the signal. The volumes were realigned and unwarped and slice-time corrected. Then, the realigned functional images were coregistered with the anatomical image and were normalized to 3 mm^3 voxels using the parameters from the segmentation of the anatomical image. Last, images were smoothed using an 8 mm Gaussian kernel, and a 128 high-pass filter was employed to remove low-frequency artefacts. Multivariate analyses used non-normalized and non-smoothed data (Bode & Haynes, 2009; Gilbert & Fung, 2018; Woolgar et al., 2011a, 2011b).

3.3.5. fMRI analyses

3.3.5.1. Univariate

First, we employed a univariate approach to examine the effect of context demands (one vs. two sequential tasks) and task switching. Our model contained, for each run, one regressor for the instruction of each miniblock, four regressors corresponding to the two types of miniblock (pure/mixed) with separate regressors for the pre-switch and post-switch periods, one for the change in fixation colour during mixed miniblocks (indicating a switch event), one for the change in fixation colour during pure miniblocks (serving as a baseline for the switch events), and another one for the errors. Both instruction and miniblock regressors were modelled as a boxcar function with the duration of the entire pre/post switch period or instruction duration (5 s). Errors were modelled including the duration of the face of that trial and the following fixation (2.5-3 s). Switch trials were modelled as events, with stick functions with zero duration locked at the switch in colour of the fixation point. This provided a model with a total of 8

regressors per run. At the group level, t-tests were carried out for comparisons related to the effect of task demands (one vs. two tasks) at the period prior to the switch, and also to compare switching cost effects (switch trial in the mixed block > switch trial in pure blocks). We report clusters surviving a family-wise error (FWE) cluster-level correction at p< .05 (from an initial uncorrected threshold $p < .001$). Additionally, we also performed nonparametric inference (see Supplementary Materials).

3.3.5.2. Multivariate analysis

We performed multivoxel pattern analyses (MVPA) to examine the brain areas maintaining the representation of A) current-active initial tasks, and B) intended tasks. These analyses examined brain activity during the pre-switch period only (although additional, exploratory analyses were also performed on the post-switch period, see the Supplementary Materials). Following a Least-Squares Separate Model approach (LSS; Turner, 2010) we modelled each miniblock (EG, ER, GE, GR, RE, RG) during the period prior to the switch separately. This method helps to reduce collinearity between regressors (Abdulrahman & Henson, 2016), by fitting the standard hemodynamic response to two regressors: one for the current event (a type of miniblock prior to the switch) and the second one for all the remaining events. As in the univariate approach, each miniblock regressor was modelled as a boxcar function with the duration of the entire pre-switch period duration.

The binary classification analyses were performed as follows. First, we classified a) between any two initial tasks while holding the intended task constant, then b) between any two intended tasks while holding the initial task constant. For instance, in case a) we contrasted initial gender task vs. initial race task when the intended task was

emotion (GE vs. RE), and also intended gender vs. intended emotion task when the intended task was race (ER vs. GR), and intended emotion task vs. intended race task when the intended task was gender (EG vs. RG). We then averaged decoding accuracies across these analyses, which indicate whether a particular brain region shows different patterns of activity depending on what the initial, currently-active task set is, holding the intended task constant. Conversely, in case b), we compared intended gender vs. intended race when the initial, currently-active task was emotion (EG vs. ER), intended gender vs. intended emotion when the initial, currently-active task was race categorization (RE vs. RG) and intended emotion vs. intended race when the initial, currently-active task was gender (GE vs. GR). As above, we averaged across these analyses, which indicate whether a particular brain region shows different patterns of activity depending on what the intended task set is, holding the currently-performed task constant.

To carry out these analyses, we performed a whole brain searchlight (Kriegeskorte, Goebel, & Bandettini, 2006) on the realigned images employing the Decoding Toolbox (TDT; Hebart, Görgen, & Haynes, 2015) and custom-written MATLAB code. We created 4-voxel radius spheres and for each sphere, a linear support vector machine classifier (C =1; Pereira, Mitchell, & Botvinick, 2009) was trained and tested using a leave-one-out cross-validation. Due to the nature of the paradigm and the counterbalancing, once in each block the switch took place at the first trial (here particcipants only performed the intended task). Thus, there was an example of each type of miniblock before the switch in only 11 runs, differently for each participant and miniblock (i.e. a participant could lack miniblock EG in run 4 and miniblock RG in run 11). To avoid potential biases in the classifier for having only one of the classes in a

run, for each participant and comparison, we performed the classification only in the 10 runs where there was an example of both miniblocks. Resulting from this procedure, we employed the data from 10 scanning runs (training was performed with data from 9 runs and tested on the remaining run, in an iterative fashion). In the exceptional case (twice for each contrast in the total sample) where the two miniblocks in the classification were absent in the same run, classification was performed on the remaining 11 runs (training with data from 10 runs and testing on the remaining run). In addition, we observed biases in the decoding estimates when the switch trial from one of the conditions in the test set matched the opposite class in the training set, which happened for every comparison in aproximately half of the cross-validation steps. To avoid the biases resulting from this, we additionally removed those runs where the switch position matched the test from the training set for that specific cross-validation step.

Next, we averaged the accuracy maps for a) and b) to obtain a mean classification map collapsing across initial and intended tasks. This allowed us to detect regions that contained information about either initial or intended tasks (or both). It also allowed us to define ROIs that could be used to compare decoding accuracies for initial versus intended task-sets, in a manner that was unbiased between the two types of information. We additionally conducted whole-brain analyses investigating decoding of the initial task only, decoding of the intended task only, and the comparisons between the two.

Afterwards, group analyses were performed by doing one-sample t-tests after normalising (same as for the univariate analyses) and smoothing the individual accuracy maps (4 mm Gaussian kernel, consistent with earlier MVPA studies such as Gilbert, 2011; Gilbert & Fung, 2018). Results were considered significant if they passed an

FWE cluster-level correction at p< .05 (based on an uncorrected forming threshold of p< .001). This statistical approach is consistent with recent MVPA studies (Gilbert & Fung, 2018; Loose, Wisniewski, Rusconi, Goschke, & Haynes, 2017). We additionally carried out nonparametric inference (see Supplementary Materials).

3.4. Results

3.4.1. Behaviour

First, to study how the number of tasks influenced performance, we performed a paired t-test on both accuracy and reaction times (RTs), between the two types of Miniblock (Mixed/Pure), collapsing over pre- and post-switch periods (see Figure 3). Here, responses were more accurate for pure (M = 95.7%, SD = 4.3), than for mixed miniblocks (M = 92.3%, SD = 3.6), $t_{31} = 5.39$, p<.001, whereas they were faster for pure (M = 671.12 ms, SD = 126.3) than for mixed (M = 709.18 ms, SD = 142.39) miniblocks, $t_{31} = 4.83$, p<.001.

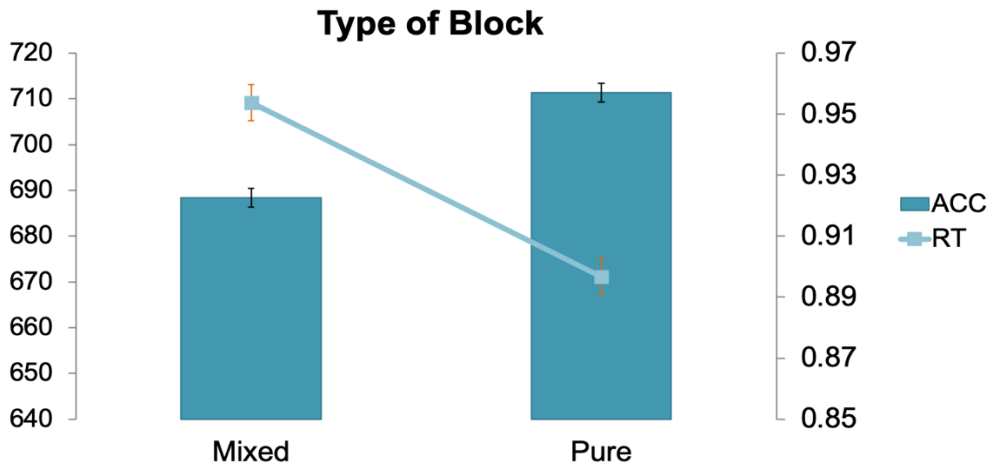

Figure 3. Influence of the type of block on performance. Error bars represent within-subjects 95% confidence intervals (Cousineau, 2005).

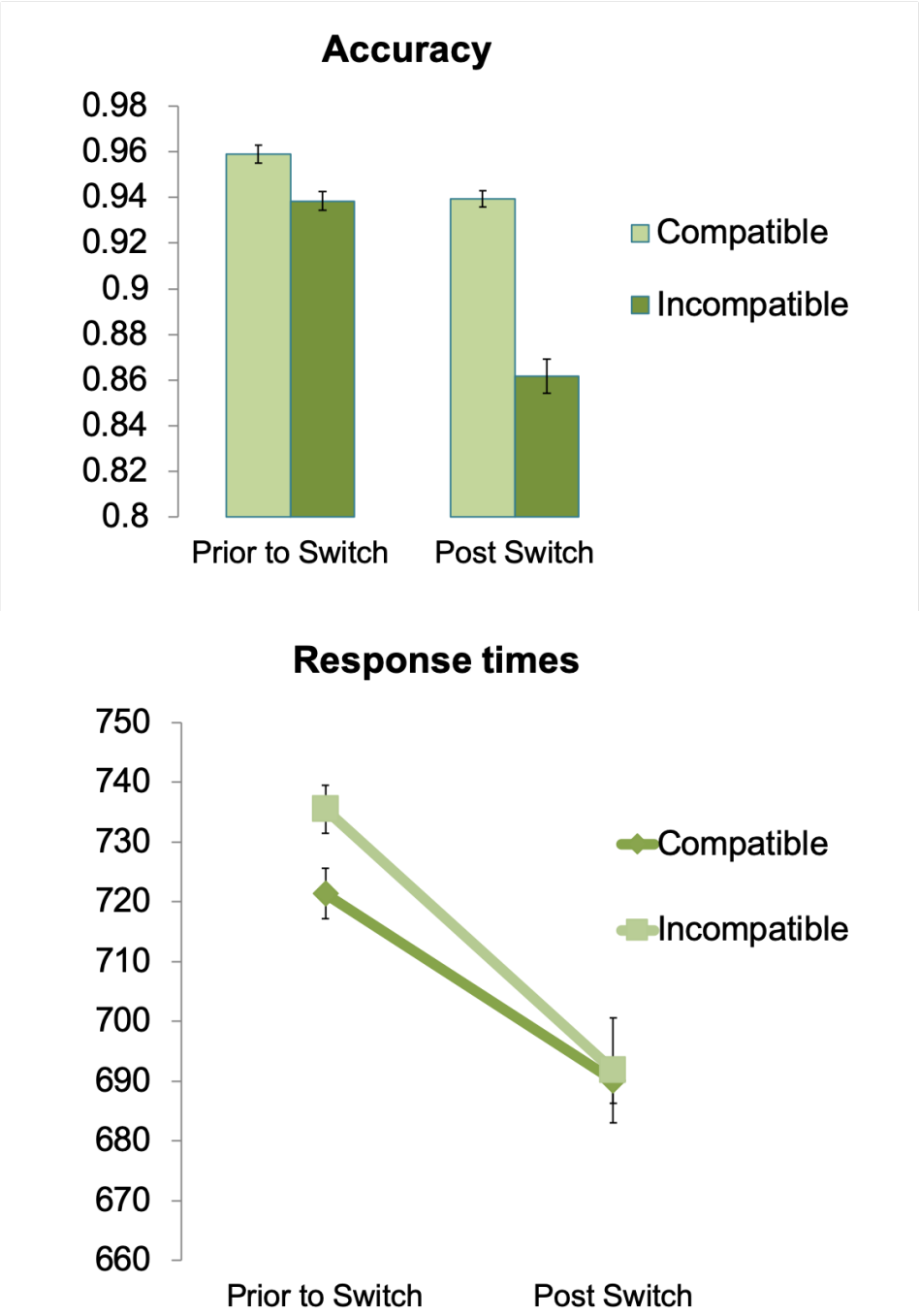

Figure 4. Interference effects between initial and intended tasks before and after the switch (MMs). Top: Accuracy rates. Bottom: Reaction times (ms). Error bars represent within-subjects 95% confidence intervals (Cousineau, 2005).

In addition, we examined if the intended task influenced initial task performance, and vice versa. For this, we selected only the mixed miniblocks and entered them into a repeated measures (rm) ANOVA, with Task (Emotion/Gender/Race), Period of the

miniblock (Prior to/Post Switch), and Interference (Compatible/Incompatible) between initial and intended tasks. Note here that even if we did not have any specific hypothesis about the influence of the variable Task on performance, we included it as a factor in this second ANOVA to examine whether the Task modulated the effect of the other two variables of interest: Period of the block and Interference. Moreover, we refer to initial and intended as the tasks performed before and after the switch, respectively, to preserve consistency in the terminology throughout the entire manuscript. In addition, we refer as compatible trials when the correct response for the initial task-relevant dimension was associated with the same response for the intended dimension (see an example in Figure 1, right), and incompatible trials when the response associated with the initial task-relevant dimension interfered with the responses associated with the intended one. Similarly, after the switch, when the intended task was being performed, compatible trials referred to those where the correct responses for this task were associated with the same response for the previous initial task, and incompatible trials when the response associated with both dimensions differed. This way we could use interference effects as an indicator of the maintenance of the intended response dimensions during performance.

Accuracy did not show any main effect of Task ($F<1$, see Figure 4). However, we observed a main effect of Period of the miniblock, $F_{1,31} = 58.215$, $p < .001$, $\eta_p^2 = .653$, where participants responded more accurately before (M = 94.86%, SD = 3.6) than after the task switch (M = 90.05%, SD = 5.4). There was also a main effect of Interference, $F_{1,31} = 101.83$, $p < .001$, $\eta_p^2 = .767$, where accuracy was higher for compatible (M = 94.91%, SD = 4.76) than for incompatible trials (M = 90%, SD = 6.79). The interaction Task x Period was significant, $F_{2,62} = 3.831$, $p = .029$, $\eta_p^2 = .110$, showing that

performance was better before than after the switch for all three tasks (all Fs > 15, ps < .001), but this difference was larger for the gender task (η_p^2 = .679). Similarly, the interaction of Period x Interference was significant, $F_{1,31}$ = 52.244, p < .001, η_p^2 = .106, where accuracy was worse for incompatible compared to compatible trials (both Fs > 19, ps < .001), but this pattern was more pronounced after ($F_{1,31}$ = 107.08, p<.001) than before the switch ($F_{1,31}$ = 19.21, p<.001). No other interactions reached significance (p>.061).

RTs showed (see Figure 4) a main effect for Task ($F_{2,62}$ = 24.08, p < .001, η_p^2 = .437), were race was performed faster (M = 691.16, SD = 150.07), followed by gender (M = 709.29, SD = 147.11), and emotion (M = 728.59, SD = 144.14). In addition, we also observed a main effect of Period ($F_{1,31}$ = 32.83, p < .001, η_p^2 = .514), as participants were faster after (M = 690.94, SD = 143.38) than prior to the switch (M = 728.42, SD = 155.72). Further, we found a main effect of Interference, $F_{1,31}$ = 4.829, p =.036), where participants were faster for compatible (M = 705.51, SD = 147.79) than incompatible trials (M = 713.65, SD = 146.42). An interaction Period x Interference ($F_{1,32}$ = 24.08, p = .033, η_p^2 = .437) showed that this interference effect was significant before the switch ($F_{1,31}$ = 8.15, p = .008), but not after ($F_{1,31}$ = .178, p>.67). None of the other interactions were significant (p>.2).

During mixed blocks we observed higher accuracies and reaction times during the period before the switch and the opposite pattern (low accuracies and faster responses) after it, which could indicate a trade-off in the data. To address this possibility, we additionally performed Pearson correlations between mean accuracy and reaction times

during both periods of the miniblock. Results show no association between the two measures, neither before (r = .07; p > .35) or after (r = .17, p > .17) the switch.

3.4.2. fMRI

3.4.2.1. Univariate

Pure vs. Mixed blocks

Before the switch, the right middle frontal gyrus (k = 56; MNI coordinates of peak voxel: 42, 44, 23) showed higher activation when participants had to maintain two tasks vs. one (Mixed > Pure blocks). Conversely, in this scenario, we observed decreased activation (Pure > Mixed blocks) in a set of regions. These included the bilateral middle cingulate cortex (k = 251; -3, -10, 33), bilateral medial prefrontal cortex (mPFC; k = 80; -6, 44, 41), left orbitofrontal cortex (OFC; k = 117; -33, 32, -16), right inferior frontal gyrus (IFG)/OFC (k = 168; 54, 32, -1), left lingual and parahippocampal gyri (k = 148; -9, 52, 5) and left superior temporal gyrus (k = 62; -57, -1, -4).

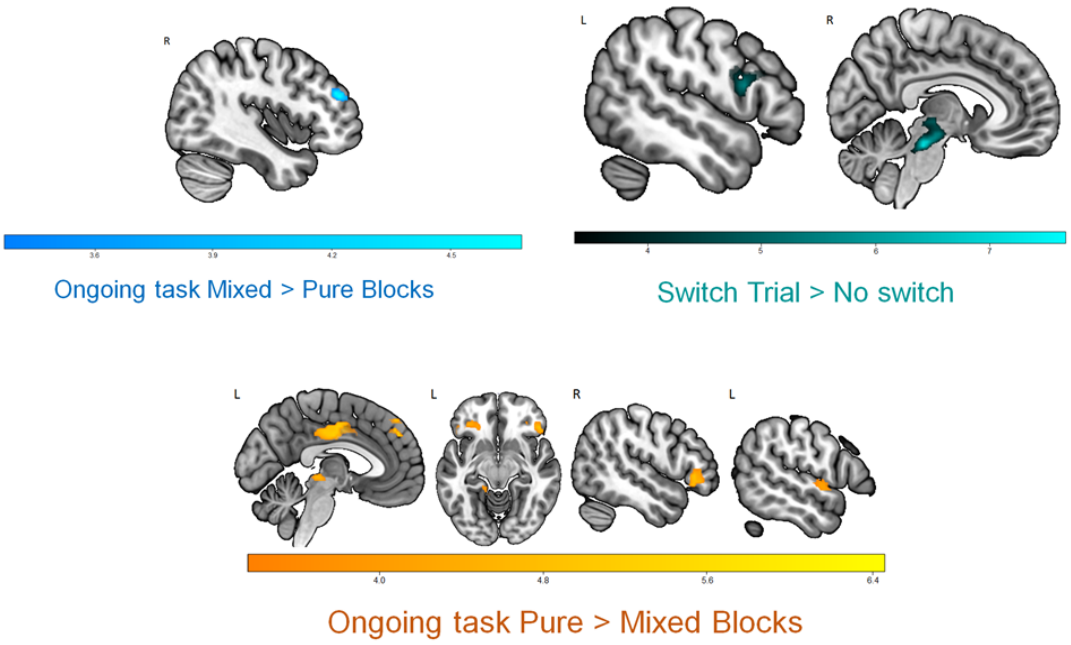

Figure 5. Univariate results. Effect of task demands (one vs. two) and task switching. Scales reflect peaks of significant t-values (p<.05, FWE-corrected for multiple comparisons).

Switch vs. non-switch trials

Transient activity during task switching was observed in the bilateral brainstem and thalamus (k = 291; -3, -28, -22) as well as in a cluster including the left inferior/middle frontal gyrus (IFG/MFG) and precentral gyrus (k = 160; -51, 11, 17).

Interference effects

We assessed whether the interference effects observed on behaviour were matched at a neural level, by comparing incompatible vs. compatible trials before and after the switch. For this model, we included in each run one regressor for the instruction of each miniblock, one regressor corresponding to all Pure Miniblocks, four regressors corresponding to Compatible/Incompatible trials in Mixed Miniblocks, with separate regressors for the pre-switch and post-switch periods (onset at face presentation) and a last regressor for errors. Instructions and errors were modelled as described in section 2.5.1 Univariate analysis. Pure Miniblocks were modelled as a boxcar function with the duration of the entire pre/post switch period, while the four remaining regressors (Compatible Pre, Compatible Post, Incompatible Pre, Incompatible Post) were modelled as events with zero duration. This provided a model with a total of 7 regressors per run.

At the group level, t-tests were carried out for comparisons related to the effect of Interference before and after the switch. No effect survived multiple comparisons. However, at a more lenient threshold ($p<.001$, uncorrected), data showed higher activation in the left IFG (k = 22; -48, 35, 20) for incompatible > compatible trials. The opposite comparison (compatible > incompatible before the switch) or incompatible vs. compatible contrasts after the switch did not yield any significant results.

3.4.2.2. Decoding results

First, we averaged all individual classification maps to examine the regions sensitive to any kind of task (initial or intended) during the period prior to the switch. We found that the rostromedial PFC/orbitofrontal cortex (OFC) presented significant accuracies above chance (k = 68; 15, 56, -10). To further examine whether one of the tasks dominated the classification, we extracted the decoding values from the initial and intended classification from the general decoding ROI above. A paired t-test between the initial and intended task decoding values showed no differences between decoding accuracy for the two types of information, $t_{31} = .846$, $p = .404$. To further test this idea, we ran an ROI analysis in this region to see whether decoding accuracy was significant for both the initial and intended task. To avoid non-independency, we employed a Leave-One-Subject-Out (LOSO) cross-validation approach (Esterman & Yantis, 2010) to select the ROI per participant. That is, data from each participant was extracted from an ROI that was defined based on the data from all the other participants, to avoid 'double dipping' (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). After this, both the initial and intended decoding values showed significant accuracy above chance (1-tailed, $t_{31} = 3.018$, p=.0025 and $t_{31} = 2.299$, $p = .0145$, respectively). This suggests that the mPFC/OFC region, prior to the switch, carries information about the relevant tasks to perform, regardless of whether they are initial or intended.

To further characterise the information represented in the mPFC/OFC region, we additionally examined whether we could cross-classify between the initial and intended tasks, which speaks to the potential overlap between these representations. Therefore, we performed ROI cross-classification analysis (Kaplan, Man, & Greening, 2015) in the mPFC/OFC region from the general decoding. We followed the same classification

procedure as described in the methods section, but training the classifier on the initial task and testing on the intended task, and vice versa. Note that in this scenario the cross-classification was carried out in a totally independent and orthogonal analysis to that employed to define the OFC region. However, results showed no significant cross-classification in this region, $t_{31} = 1$, $p > .3$, which suggests that the initial vs. intended nature of the tasks may change their representational format.

Moreover, we examined the initial and intended individual classification maps separately to examine the regions sensitive to each type of task (initial or intended). With this, the classification of the initial task alone showed a different cluster in the left OFC (k = 52, -15, 17, -13). Conversely, we observed the right IFG (k = 53, 45, 41, -16) for the classification of the intended task alone. Last, when comparing decoding accuracies between the initial and intended tasks (subtracting initial – intended accuracy maps), we observed significantly higher accuracies for the representation of the initial (vs. intended) in the right FG and the hippocampus (k = 148; 39, -13, -25) and also in left OFC (k = 44; -18, 17, -16). However, the opposite contrast (intended vs. initial) did not show any cluster with significant differences.

Importantly, these MVPA results were unlikely to be due to differences in response times (see Supplementary Materials *"RT analyses between miniblock pairs and correlation with decoding accuracies"*).

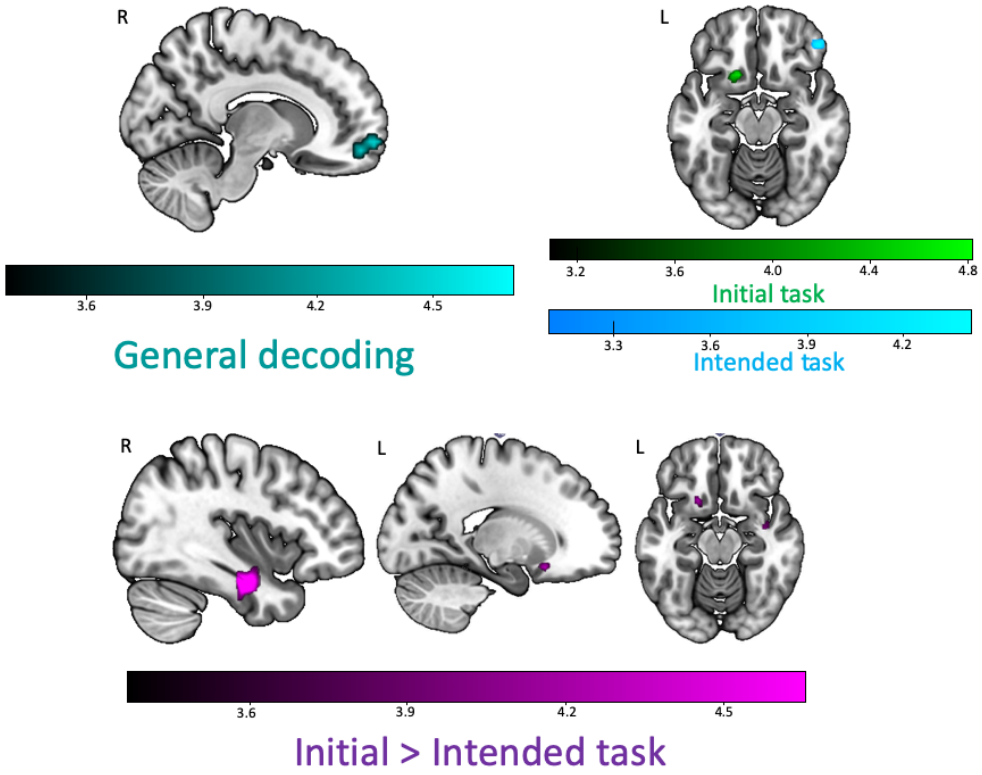

Figure 6. Multivariate results during the period prior to the switch. Top left: General decoding (cyan) of the region sensitive to any kind of task (initial or intended). Top right: Decoding of the initial (green) and intended (blue) task separately. Bottom: Decoding of initial > intended task (violet). Scales reflect peaks of significant t-values ($p<.05$, FWE-corrected for multiple comparisons).

Last, although the main focus of this work is the period before the switch, we performed exploratory classification analysis in the period after the switch to obtain further understanding about how information is represented once the initial task is no longer active. Following the same procedure as in section *2.5.2 Multivariate analysis*, we carried out the decoding analysis with regressors of interest corresponding to the period after the switch. As in the previous analyses, we first averaged all individual classification maps to examine the regions sensitive to any kind of task during the period after the switch. Here, we observed a cluster in left middle frontal/precental gyrus (k = 61; -36, 11, 35) with significant accuracies above chance. Next, we averaged

the classification maps separately for the currently-active post-switch task and the previously-performed initial task, to examine the regions sensitive to each type of task. The initial task could be decoded from bilateral IFG (left; *k* = 48; -54, 29, 26/right; *k* = 45; 54, 32, 5). Conversely, decoding of the currently-active post-switch task could only be observed at a more liberal threshold ($p < .01$, k = 100) in a cluster including the FG/cerebellum (*k* = 109; 15, -52, -19) and in left middle/inferior frontal gyrus (*k* = 122; -45, 8, 35).

3.5. Discussion

The present work aimed to examine the representation of A) currently-active, initial task sets, and B) intended task sets applied to social stimuli during a dual-sequential social categorization of faces. We found some regions from which we could decode only the currently-active or the intended task set, along with a region of vmPFC/OFC containing both types of information although cross-classification between the two was not possible.

The paradigm employed revealed behavioural costs due to the maintenance of intended task sets. Here, mixed blocks presented slower responses and lower accuracies, compared to pure ones, in line with previous studies (Los, 1996; Marí-Beffa et al., 2012). Response times were slower before than after the switch, when the intended task needed to be held while performing the initial one, in line with Smith (2003). These results reflect the higher demands associated with the maintenance of two relevant task sets. In addition, people need more time to respond when they hold the intention while performing the initial task, but not after, when they only need to focus on a single task. Further, we aimed to examine the activation of both initial and intended task sets before

the switch by looking at possible interference between them. Behavioural results showed incompatibility effects in performance when participants needed to hold the intended task set while performing the ongoing task. These results point to the maintenance of intended task settings, showing that information about the pending set is maintained online during performance of the ongoing task.

Turning to brain activation data, the behavioural costs we observed during the maintenance of two task sets during mixed blocks were accompanied by increased activation in the right MFG, while regions such as the left IFG and the thalamus increased their activation during switching trials. Incompatible trials during this period were also related to activation in left lPFC, but only at an uncorrected statistical threshold. These results fit with previous work that associated lPFC with sustained control during dual tasks (Braver et al., 2003; Szameitat et al., 2002) and the maintenance of delayed intentions (Gilbert, 2011). Lateral PFC has also an important role in rule representation and the selection of correct rules (Brass & Cramon, 2002; Crone, Wendelken, Donohue, & Bunge, 2006), while the thalamus has also shown a role in cognitive flexibility during task switching (Rikhye, Gilra, & Halassa, 2018). Overall, this pattern reflects the cognitive demands of holding two tasks in mind, extending the results to the maintenance of social categorization sets.

In addition, we observed a set of task-negative regions such as mPFC, middle cingulate, OFC, IFG and temporal cortex which showed reduced signal when participants needed to hold an intended task set. Similarly, Landsiedel & Gilbert (2015) carried out an intention-offloading paradigm where participants had to remember a delayed intention, which they had to fulfil after a brief filled delay. During the maintenance of such

intention they found a set of deactivated regions including mPFC, posterior cingulate cortex, infero-temporal cortex, and temporo-parietal cortex. Interestingly, these authors showed that when participants had the opportunity to offload intentions by setting an external reminder, this decrease of activation was ameliorated. This reduction suggests that these task-negative areas play a role in the representation of the delayed tasks. Pure blocks in our tasks are similar to the offloading condition in Landsiedel & Gilbert (2015), as participants needed to sustain only the initial task set to perform the correct social categorization. Therefore, taking these results into account, it seems unlikely that task-negative regions reflect simply a "default mode" (Fox et al., 2005), but rather that this deactivation is, to some extent, playing a functional role during task performance (Spreng, 2012).

Currently, multivariate pattern analyses are one of the most powerful approaches to study the information contained in different brain areas. In this work, we employed MVPA to study the regions representing initial and intended task sets. Importantly, a novelty of our approach was to do so by employing three different tasks instead of just two, unlike most previous studies (e.g. Haynes et al., 2007; Momennejad & Haynes, 2013). We were able to distinguish different regions representing initial and intended task sets. The currently-active initial task was decoded from left OFC, while information about the intended one was contained in the right IFG. Also, comparing between these two, a nearby OFC region together with the FG showed higher fidelity of the representation for the initial vs. intended task set. The left OFC region found for the initial task decoding has been previously associated with representations of facial information related to social categories (Freeman, Rule, Adams, & Ambady, 2010) and it also facilitates object recognition in lower level areas such as the FG (Bar et al.,

2006). In addition, the FG showed greater fidelity of the representation of initial vs. intended task sets, which could indicate the allocation of attentional resources to process current relevant information in earlier perceptual regions. Besides having a prominent role in processing faces (Haxby et al., 2000), previous studies have been able to decode task-relevant information related to social categories (e.g. female vs. male, black vs. white) in perceptual regions such as the FG (Gilbert et al., 2012; Kaul et al., 2014, 2011; Stolier & Freeman, 2017). On top of this, previous work using facial stimuli in the field of social perception and prejudice has shown how the FG is influenced by stereotypical associations and evaluations of social categories associated with the OFC (Gilbert et al., 2012; Stolier & Freeman, 2016). Our data fit and extend this work, suggesting that both OFC and FG contain information about the appropriate task set to perform the initial classification, pointing out the role of both high and lower-level perceptual regions in the representation of ongoing task sets performed on facial stimuli.

Conversely, decoding of the intended task was possible from the IFG. Previous work has related lPFC with the representation of intentions (Haynes et al., 2007; Momennejad & Haynes, 2012; Momennejad & Haynes, 2013; Soon et al., 2008) and it has also been linked to task-set preparation (e.g. González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; González-García, Mas, de Diego-Balaguer & Ruz, 2016; Sakai & Passingham, 2003). This result, however, contradicts the proposal of Gilbert (2011) in which lPFC would serve as a general store for delayed intentions without information about their content. Nonetheless, while Gilbert (2011) decoded simpler stimulus-response mappings, in our study we classified abstract task-set information, as previous studies that also decoded intention from lPFC did (e.g. Haynes et al., 2007; Momennejad & Haynes, 2013). Therefore, our work agrees with Momennejad & Haynes (2013),

suggesting that lPFC may represent delayed intentions when their content is abstract enough.

Interestingly, we found a region in vmPFC/OFC that contained information about both initial and intended task sets. We examined if this vmPFC/OFC region contained overlapping representations of both initial and intended task sets by performing a cross-classification analysis, and observed a lack of generalization from the two sets of representations. Our cluster is close to the mPFC region found by previous intention studies (e.g. Gilbert, 2011), but located in a more ventral area. Notably, in our paradigm both the initial and intended tasks that participant performed were equally relevant and were based on the same set of stimuli. Altogether, this suggests that mPFC may be recruited when relevant task sets need to be held for a period of time, irrespective if this information is being using at the moment or later. This result also fits with a recent proposal (Schuck, Cai, Wilson, & Niv, 2016) characterising this region as a repository for cognitive maps of task states. Specifically, the vmPFC/OFC would represent task-relevant states that are hard to discriminate based on sensory information alone (Schuck et al., 2016; Wilson, Takahashi, Schoenbaum, & Niv, 2014), similar to our study. Overall, this suggests a role for this region in the maintenance of task-relevant information, regardless of when it is needed. Given that cross-classification between initial and intended task sets was not possible in this region, this would be compatible with multiplexed representations of the two types of information (i.e. representations using orthogonal representational codes). It could also be compatible with nonoverlapping populations of cells representing initial and intended tasks within the voxels in this region. Further, it could be argued that the lack of cross-classification between initial and intended tasks could be explained by differential processes

underlying the representation of these tasks. That is, the initial one could be represented simply by the activation of the S-R associations, while the intended one would rely on the maintenance of a general task set. Nonetheless, an explanation relying solely on this distinction seems unlikely, since we observed interference effects. Even though the interference effect for accuracy was smaller before the switch, for reaction times we observed interference only during this period. This indicates that the S-R association for both tasks is active prior the switch, although there could be some distinctions in the way these response representations are maintained. However, the lack of significant cross-classification could also reflect simply a lack of statistical power.

Despite our decoding results, we did not find the dissociation pattern that we predicted for initial and intended task sets. Although we could decode intended task sets from mPFC, consistent with previous studies (Gilbert, 2011; Haynes et al., 2007; Momennejad & Haynes, 2012; Momennejad & Haynes, 2013), we did not find a frontoparietal representation of initial task sets as previous studies have (Qiao et al., 2017; Waskom et al., 2014; Woolgar et al., 2011b). This discrepancy could be explained by differences in the stimuli and task employed. Previous studies have decoded task information in the form of classification between different stimulus-responses mappings or different types of stimuli. In contrast, in our task we employed the same stimuli for all tasks, and the differences between the information to decode relied in the representation of the social category needed for the specific part of the miniblock, rather than perceptual features (e.g. the colour or shape of target stimuli), which may be easier to decode (Bhandari, Gagne, & Badre, 2018).

To conclude, in the present work we examined the representation of A) currently-active initial task sets, and B) intended task sets during a social categorization dual-sequential task. Crucially, we directly examined the common and differential representation of initial and delayed task sets, extending previous work studying these mechanisms separately. Moreover, we employed faces as target stimuli to complement prior research. Apart from replicating previous findings in dual tasks with social stimuli, we show how task set information was contained in different regions, depending on when it was needed. Thus, currently-active initial tasks were represented in specialized regions related to face processing and social categorization such as the OFC and FG, while intended ones were represented in lPFC. On top of that, we showed a common brain region in vmPFC/OFC maintaining a general representation of task-relevant information, irrespective of when the task is performed, albeit it is not clear whether overlapping patterns of activation represent both types of information. Moreover, the results from the classification after the switch suggest that the representation of the two relevant tasks varies once the switch takes place and the initial task is no longer active. Future research should. further characterize the representational format of relevant task information depending on when it is needed and examine the structure of the task set representation within these regions, for instance using Representational Similarity Analysis (RSA, Kriegeskorte, Mur, & Bandettini, 2008). Also, studying the interaction of sustained task sets with specific conditions of each trial would extend our knowledge of the representational dynamics of current and intended task-relevant information. Last, due to the social significance of faces, one step forward would be to examine how their representation may vary in social scenarios, contexts in which faces and other social stimuli are particularly relevant to guide behaviour (Díaz-Gutiérrez, Alguacil, & Ruz, 2017).

3.6. Supplementary Material

3.6.1. fMRI non-parametric analyses

3.6.1.1. Univariate

Permutations tests were carried out employing statistical non-parametric mapping (SnPM13, http://warwick.ac.uk/snpm) with 5000 permutations. We performed cluster-wise inference on the resulting voxels with a cluster-forming threshold at 0.001, which was later used to obtain significant clusters (FWE corrected at $p<0.05$).

3.6.1.2. Decoding

For the MVPA, we followed the method proposed by Stelzer, Chen, & Turner (2013), especially suitable for MVPA data. As in Arco, González-García, Díaz-Gutiérrez, Ramírez, & Ruz (2018), we permuted the labels and trained the classifier 100 times for each participant. After normalizing the resulting maps to an MNI space, for each participant, we randomly picked one of these maps, averaged them and obtained a map of group permuted accuracies. This method was repeated 50000 times to build an empirical chance distribution for each voxel. The 50th greatest value corresponds to the threshold for statistical significance at $p<0.05$, false-discovery rate (FDR) corrected.

3.6.2. fMRI non-parametric results

3.6.2.1. Univariate

Pure vs. Mixed blocks

Before the switch, the right middle frontal gyrus (k = 77; 42, 44, 23) showed higher activation when participants had to maintain two tasks versus one. Conversely, in this scenario, we observed decreased activation in a set of regions. Here, we observed cluster deactivation in bilateral middle cingulate cortex (k = 274; -3, -10, 33), bilateral medial prefrontal cortex (mPFC; k = 86; -6, 44, 41), left orbitofrontal cortex (OFC; k = 148; -33, 32, -16), right inferior frontal gyrus (IFG)/OFC (k = 193; 54, 32, -1), right parahippocampal/fusiform gyri and hippocampus (k = 37; 27, -19, -22), left

parahippocampal/fusiform/inferior temporal gyri (k = 49; -21, -19, -28) and left lingual and parahippocampal gyri (k = 257; -9, 52, 5).

Switch vs. non-switch trials

Transient activity during switching tasks was observed in the bilateral brainstem and thalamus (k = 380; -3, -28, -22) as well as in a cluster including the left IFG/MFG and precentral gyrus (k = 210; -51, 11, 17). Further, these trials increased activation bilaterally in the anterior insula/IFG (k = 31; -24, 23, -7/k = 30; 24, 26, -4).

Incompatibility effects

Last, we assessed whether the compatibility effects observed on behaviour were matched at a neural level, by comparing incompatible vs. compatible trials before and after the switch. Data showed higher activation in the left IFG (k = 42; -48, 35, 20) for incompatible > compatible trials before the switch. The opposite comparison (compatible > incompatible before the switch) or incompatible vs. compatible contrasts after the switch did not yield any significant results.

3.6.2.2. Decoding

First, we averaged all individual classification maps to examine the regions sensitive to any kind of task (initial or intended) during the period prior to the switch. Here, we found that the rostromedial PFC/orbitofrontal cortex (OFC) presented significant accuracies above chance (k = 48; 15, 56, -10). Further, looking at the classification of the initial task only, we observed a cluster in the right OFC (k = 13, 15, 56, -13). However, in the classification for the intended task, we did not find any region showing significant accuracies above chance.

Last, when comparing decoding accuracies between the initial and intended tasks (subtracting initial – intended accuracy maps), we observed significantly higher

accuracies for the representation of the initial (vs. intended) tasks in a cluster (k = 24; 39, -13, -25) covering the right FG and the hippocampus. Nonetheless, the opposite contrast (intended vs. initial) did not show any cluster with significant accuracies above chance.

3.6.3. RT analyses between miniblock pairs and correlation with decoding accuracies

In order to assess the influence of RT difference between tasks in decoding accuracies, we carried out a comparison between miniblock pairs. In line with the classification analyses, we performed paired t-tests between these pairs during the period before the switch: EG vs. RG/ER vs. GR/GE vs. RE for initial task differences and EG vs. ER/GE vs. GR/RG vs. RE for the intended task. Doing this, two contrasts of the initial tasks were significant: emotion vs. race (EG vs. RG, p=.003) and emotion vs. gender (ER vs. GR, p =.008), but no differences for gender vs. race (GE vs. RE, p>.1) or any of the comparisons for the intended tasks (all ps>.3).

Then, similarly to Crittenden et al. (2015) we obtained absolute RT differences for the initial task and intended task discriminations and conducted a Spearman's correlational analysis of RTs against decoding accuracies for each result ROI. As can be seen in Table 1, results show that decoding is not related to response times' differences between the conditions that have been classified.

Supplementary Table 1. Spearman's correlational analysis of RTs against decoding accuracies.

	Spearman r	p-value
Relationship between the general decoding ROI and response times.		
general RT differences and general decoding	-.181	.134
initial RT differences and general decoding	-.201	.322
intended RT differences and general decoding	-.141	.442
initial RT differences and general decoding, initial	-.196	.442
intended RT differences and general decoding, intended	-.252	.164
Relationship between the initial decoding ROI and response times		
initial RT differences and OFC decoding	.008	.965
Relationship between the initial > intended decoding ROIs and response times		
initial RT differences and OFC decoding	.087	.636
initial RT differences and FG decoding	.025	.892
Relationship between the intended decoding ROI and response times.		
intended RT differences and IFG decoding	.152	.407

CHAPTER 4

EXPERIMENT II

The content of this chapter is in preparation as Díaz-Gutiérrez, P., Alguacil, S. & Ruz, M. Control and interference with social and non-social information during a Trust Game.

4.1. Abstract

During social decisions we are exposed to multiple sources of information that can bias our supposedly rational choices. This makes crucial the understanding of how control mechanisms help us navigate these situations. Previous work has examined how control mechanisms adjust our behaviour at different timescales (Dosenbach et al., 2008) during complex paradigms (Palenciano et al., 2019a,b). However, it has not been studied yet the role of these mechanisms in interpersonal decisions. Thus, in this study we carried out a hybrid fMRI experiment to examine sustained and transient control mechanisms during an interpersonal Trust Game. Here, the outcome from these interactions was predicted by social or non-social dimensions of facial stimuli. Our results show general conflict markers in behaviour and transient activation of frontoparietal regions when social dimensions interfered with each other. Overall, we extend the role of control mechanisms and show transient adjustments during interpersonal decisions.

4.2. Introduction

An essential component of human behaviour is the navigation of social scenarios, where we constantly interact with others. Although this might sound easy, in these contexts we need to ponder different sources of information and predict people's likely behaviour. Here, our supposedly rational decisions are influenced by several factors that predispose us to act in a certain manner, which is not necessarily the most efficient one (Díaz-Gutiérrez et al., 2017). This makes essential the understanding of how control mechanisms arbitrate among different action tendencies triggered by different sources, to make optimal decisions. Previous work has studied how control processes adjust our behaviour at different timescales employing a variety of tasks and stimuli, yet most of them have examined these in tasks that lack a social component (Braver et al., 2003; Palenciano et al., 2019a,b). Moreover, although some studies have tried to discern how control may be implemented with different types of information (e.g. emotional vs. non-emotional conflict; Egner et al., 2008; Torres-Quesada, Korb, Funes, Lupiáñez & Egner, 2014), it has not been examined yet whether control mechanisms are similar during complex social scenarios and whether they can vary depending on the type of information that is relevant to guide optimal decisions. In this fMRI study, we aimed to examine the regulation of behaviour under the influence of social and non-social information at both sustained and transient timescales during an interpersonal Trust Game.

The ability to regulate and coordinate our behaviour according to goals is due to control mechanisms that act with different temporal profiles (Braver,

2012). First, proactive control prepares us in advance, maintaining relevant information for subsequent task performance (*task set*; Sakai, 2008). Here, lateral prefrontal cortex (lPFC) has a prominent role in task set representation, together with anterior PFC, pre-supplementary motor area (pre-SMA) and parietal cortex (Brass & von Cramon, 2004; De Baene & Brass, 2014; Palenciano et al., 2017). Conversely, reactive control responds adaptively to task demands, for instance when we face competing action tendencies that lead to conflict (Botvinick et al., 2001). The anterior cingulate cortex (ACC), specially its dorsal portion, has been associated with the detection and monitorization of this conflict (Botvinick et al., 2001), triggering adjustments implemented by the DLPFC to enhance the processing of task-relevant information in specialized areas and the inhibition of irrelevant responses (Egner & Hirsch, 2005; Hampshire, Chamberlain, Monti, Duncan, & Owen, 2010). Overall, the existing work agrees on the crucial role of a set of frontoparietal (FP) regions, also known as the Multiple Demand Network (MDN, Duncan, 2010) for their adaptive representation of task-relevant information (Woolgar et al., 2011a; Woolgar et al., 2016).

On top of that, Dosenbach et al. (2008) proposed a subdivision of control mechanisms into two components acting at different timescales, with complementary roles. First, a cingulo-opercular network including the anterior cingulate cortex (ACC), anterior insula/frontal operculum (aI/fO), and anterior prefrontal cortex would be in charge of the stable maintenance of task-sets during task blocks. Conversely, a frontoparietal network comprising the dorsolateral prefrontal cortex (DLPFC) and intraparietal sulcus (IPS) would be associated with transient adaptative control

(Dosenbach et al., 2007). This distinction has been supported by studies in non-social contexts (e.g. Crittenden, Mitchell, & Duncan, 2016). They observed, during a task-switching paradigm, that task rules were more strongly represented in FP regions. Moreover, Palenciano et al. (2019b) examined the role of these two networks during the implementation of novel instructions and they observed that the two networks acted at both sustained and phasic timescales, reflecting the adaptability of these mechanisms to specific task demands.

Despite this extensive work, it has not been examined yet the extent to which these findings generalize to social contexts. When we interact with others we are exposed to different types of information about them that we use to infer their thoughts or predict their likely behaviour (Frith & Frith, 2006). Faces are one of the most important sources of social knowledge since they convey a lot of information about people, such as their identity or emotions. People are able to make social judgements based on faces, even after a brief exposure (Todorov, Mende-Siedlecki, & Dotsch, 2013). For instance, people attribute intentions and personal traits, such as trustworthiness, to facial expressions (Todorov, Said, Engell, & Oosterhof, 2008). This way, positive emotions are associated to trust and negative ones with unreliability. These judgements have an automatic impact on decisions, even when they are not predictive of people's behaviour (Alguacil et al., 2015; Tortosa, Lupiáñez, & Ruz, 2013). This way, participants' decisions are slower when their identity-based expectations conflict with their partners' emotion or when emotions do not lead to their natural consequences (Alguacil et al., 2015; Ruz & Tudela, 2011).

In neural terms, the resolution of conflict triggered by emotional stimuli has been associated with the rostral ACC (Egner et al., 2008; Etkin et al., 2006), as well as to the aI, orbitofrontal cortex (OFC) and inferior frontal gyrus (IFG; Levens & Phelps, 2010). During social decisions, Ruz and Tudela (2011) observed that the prefrontal cortex, dACC and aI/FO increased their activation when the partners' emotions did not predict their natural consequences, while the opposite scenario was associated with activation in the precuneus. What is more, the ACC activation was coupled with the preSMA and middle cingulate cortex (MCC) when expectations were violated but increased connectivity was observed with the ventromedial PFC and precuneus (associated to mentalizing processes; Saxe, 2006) when participants could trust the partners' emotions. Similarly, Alguacil et al., (*unpublished)* employed a Trust Game to study the automatic effect of emotions where partners' identity served as a cue to guide decisions and their emotion needed to be ignored. Here, the congruence between identity and emotion increased activation in regions related to face processing, such as the lingual gyrus (LG), cuneus and fusiform Gyrus (Haxby et al., 2002; Henson et al., 2003; Zhen, Fang, & Liu, 2013), whereas incongruence engaged the dACC and medial frontal cortex. Altogether, this evidence highlights how control-mechanisms participate when social dimensions lead to opposite predictions, whereas congruence between them involves regions associated with face processing or mentalizing.

Although previous work has studied interference between emotion and identity during social decisions (e.g. Alguacil et al., 2015; Alguacil, Madrid, Espín & Ruz, 2017; Ruz and Tudela, 2011; Tortosa et al., 2013), it is not clear

yet the neural mechanisms that underlie the stable maintenance of task-relevant information of social nature during interpersonal scenarios. In this work we aimed to examine the role of the dual model of control by Dosenbach et al. (2008) during social decisions. To introduce an interpersonal scenario, we employed a Trust Game (TG; Camerer, 2003), which allows to manipulate the relevant social dimension that carries relevant information about the behaviour of partners. We employed an fMRI mixed design, manipulating the relevant dimension to guide decisions: either social (emotion, identity) or non-social (colour). This allowed us to investigate the neural regions underlying sustained activation related to specific task-sets, and phasic response to interference between different dimensions when they led to opposite predictions.

4.3. Methods

4.3.1. Participants

Thirty-eight volunteers were recruited from the University of Granada (M = 22.7, SD = 3.27, 12 men). All of them were right-handed with normal or corrected-to-normal vision. Two participants were excluded from the analyses due to excessive head movement (>3 mm). Sample size was chosen following recommendations for mixed designs (Petersen and Dubis, 2012). Participants signed a consent form approved by the Ethics Committee of the University of Granada. In exchange for their time, participants received economic remuneration (20-25 Eur, depending on performance).

4.3.2. Stimuli

Eight faces (four identities, two females) depicting happiness or anger were selected from the NimStim database (Tottenham et al., 2009). Photographs were framed in various colours, in two different tones each: dark and light yellow, orange, green and blue. Stimuli were controlled and presented by E-Prime software (Schneider et al., 2002). Inside the scanner, the task was projected onto a screen, visible through a set of mirrors placed on the radiofrequency coil.

4.3.3. Procedure and design

We employed a mixed blocked/event-related design to extract both sustained and phasic activity (Petersen & Dubis, 2012; Visscher et al., 2003). Participants played multiple rounds of a modified Trust Game with 4 different partners. In this scenario, participants received each time 2 EUR and they had to decide whether to share it or not with the partner. If participants decided to cooperate, the money was multiplied by 5 and passed to the partner, who would receive 10 EUR. If partners reciprocated, each member of the interaction could earn 5 EUR. Conversely, if the partners did not reciprocate, the participant did not receive any money on that trial. Finally, participants could also decide not to share the initial sum, in which case no money would pass to their partners. The dimension used to predict the outcome of each interaction was manipulated between blocks: two social (the identity of the partner or their facial expression) and one non-social (the colour of the frame). Whereas happy emotions were always predictive of cooperation, with the opposite being true for angry expressions, the association between identity and colour with cooperation tendencies was

counterbalanced across participants. All these cues were valid on 100% of the trials. Further, the task cue could be congruent with one or both of the other two irrelevant cues in that specific block. There was an equal number of congruent and incongruent trials, as well as the same proportion of congruency (50%) across blocks.

Participants were told that they would be interacting with game partners that represented the identities of real participants of a previous similar experiment, so the information regarding their behaviour would be based on their actual performance. In addition, participants were encouraged to play their best to obtain the largest economic rewards possible.

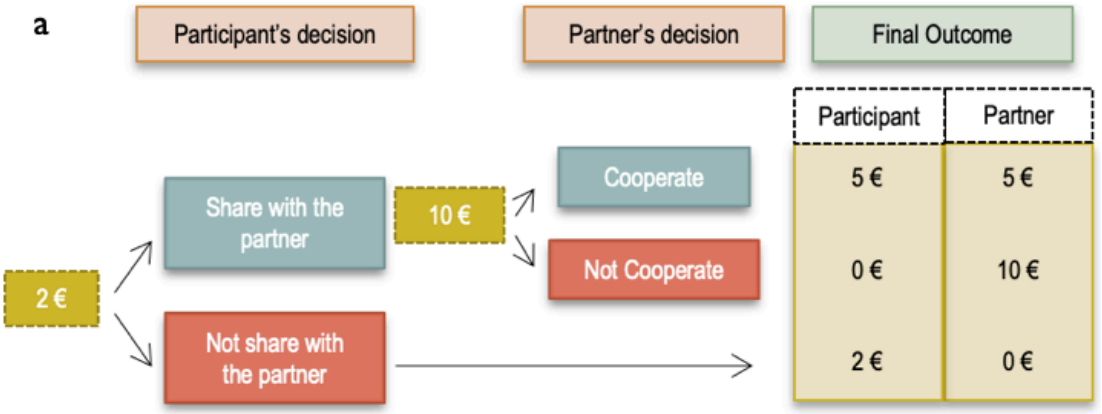

Figure 1. (a) The trust game played by participants. (**b)** Cues employed in the experiment. *Left:* Different identities associated with cooperative or non-cooperative behaviour displaying either happy or angry expressions. *Right:* Warm and cold colours used to frame the facial displays.

Prior to the scanning phase, participants were trained with a practice task where they learned about the association between the 3 dimensions and cooperation tendencies (16 trials of each block). During fMRI scanning, participants performed a total of 12 blocks, arranged in 4 runs. In each block, participants performed each task once, in a counterbalanced order. Blocks were preceded and followed by 54.7 s rest periods (+\0.5°). In each block, a 0.5s task cue preceded a variable interval (random 3.315, 5.525 or 7.735 sec), which was followed by 32 trials. Each trial started with a framed face display for 0.5 sec (7.6°) to which participants had to give their response (until 1.5 sec after face onset), followed by a jittered interval lasting 5.025 s on average (1.71, 3.92, 6.13, or 8.34 s, +\0.5°). Participants performed a total of 384 trials (96 per run) for 53.35 min.

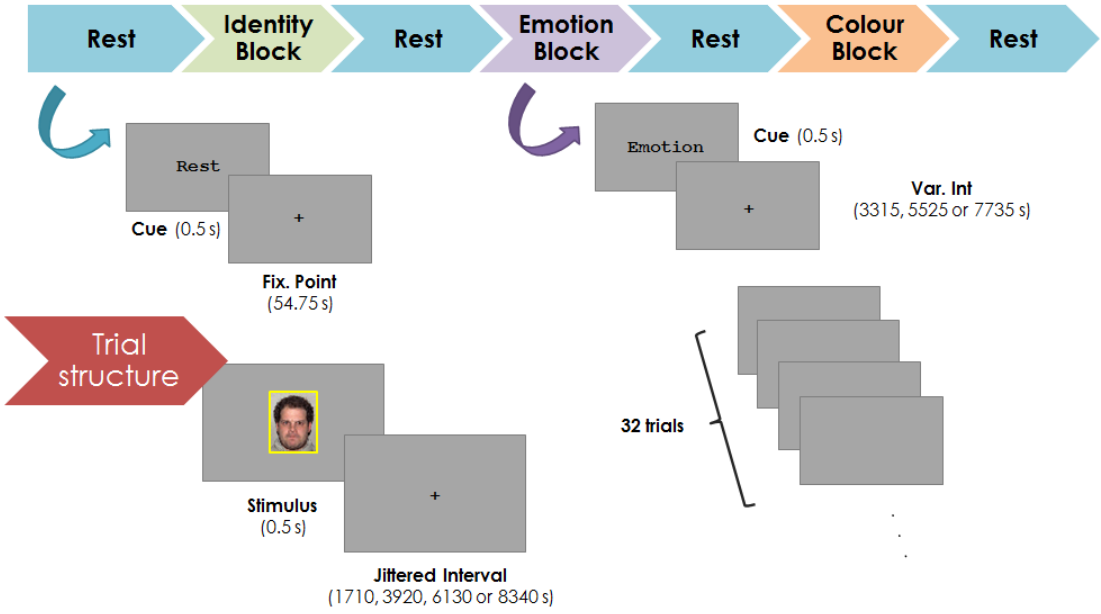

Figure 2. Overview of the experimental design (example of one run).

4.3.4. Image acquisition

MRI images were obtained using a Siemens Magnetom TrioTim 3T scanner, located at the Mind, Brain and Behaviour Research Centre in Granada. Functional images were obtained with T2*-weighted echo-planar imaging (EPI) sequence, with a TR of 2.21 s. Forty descendent slices with a thickness of 2.3 mm (20% gap) were extracted (TE = 23 ms, flip angle = 70 º, voxel size of 3x3x2.3 mm). The sequence was divided into 4 runs, consisting of 367 volumes each. After the functional sessions, a structural image was taken of each participant, using a high-resolution T1-weighted sequence (TR = 2500 ms; TE = 3.69 ms; flip angle = 7º, voxel size of 1 mm^3).

4.3.5. Preprocessing and analysis

Data were preprocessed and analyzed with SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12). The first 3 images of

each run were discarded to allow the signal to stabilize. Images were realigned and unwarped to correct for head motion, followed by slice-timing correction. Afterwards, T1 images were coregistered with the realigned functional images. Functional images were then spatially normalized according to the standard Montreal Neurological Institute (MNI) template and smoothed employing an 8 mm Gaussian kernel.

First-level analyses were conducted for each participant, following a General Linear Model. Sustained activity was modelled using boxcar-shaped regressors for each block, all convolved with the canonical hemodynamic response function (HRF), whereas transient activity was modelled with a Finite Impulse Response (FIR) basis set (9 time bins of 2.21 s), for each of the 16 event regressors (congruent and incongruent trials between the relevant and irrelevant dimensions in each block type). Rest periods and errors were also modelled as boxcar functions with their entire duration, both convolved with the HRF. At this point, t contrasts were conducted to compare different blocks (versus baseline and with each other), as well as to test how events associated with different levels of congruency and relevant dimensions differed. At the group level, t-tests were conducted to examine sustained activity (blocks against rest). For transient activity, the beta maps from the event contrasts were entered into two different full factorial ANOVAs. This approach was previously followed by Palenciano et al. (2019b) to ease contrast specification in complex designs like this one, although they replicated their results with a flexible factorial ANOVA. The first ANOVA explored general differences between tasks and interference between relevant and irrelevant dimensions. Thus, it included Task (Identity, Emotion,

Colour), Interference between the relevant block dimension and the other two (Congruent, Incongruent) and Time (9 time bins). In addition, to examine the specific interaction between the two social dimensions and their interference effects, we carried out a second ANOVA only with the social tasks. Here we included Task (Identity, Emotion), Interference between the relevant social dimension and the irrelevant social one in each block (Congruent, Incongruent) and Time (9 time bins). Note that this information could not be extracted from the previous ANOVA because it only considered trials where the relevant dimension was congruent or not with both the irrelevant dimensions, but not those where the relevant one was congruent/incongruent with only one of the irrelevant dimensions. Results were $p<.05$ FWE cluster-corrected, obtained from an initial uncorrected $p<.001$ threshold.

4.4. Results

4.4.1. Behavioural data

Correct responses (hits) and response times (RTs) of all blocks were analysed in a repeated-measures (rm) ANOVA, with Dimension (identity, emotion, and colour) and Interference with the other cues (congruent/incongruent) as factors. The Greenhouse-Geisser correction was employed when sphericity was violated. Note that as a dependent variable we employed correct responses instead of cooperation rates, since we were interested in how well participants made the most efficient decision according to their goals.

For *accuracy*, data showed a main effect of Interference, $F_{1,35} = 5.867$, $p = .021$, $\eta_p^2 = .144$, where correct responses were more frequent for congruent

(M = 96.68%, SD = 4.94) than for incongruent (M = 95.67%, SD = 6.31) trials. No other main effects or interactions were found (all p_s>.1). In *response times,* we observed a main effect of Dimension, $F_{2,34}$ = 19.038, p <.001, η_p^2 = .515, where participants were faster for the identity task (M = 740.48 ms, SD = 102.25) than for the emotion (M = 778.62 ms, SD = 116.57) and the colour (M = 785.1 ms, SD = 115.43) ones. Planned comparisons revealed that these differences were significant between identity and the other two tasks (p_s<.05) but they were not between emotion and colour (p = 1). There was also a main effect of Interference, $F_{1,35}$ = 14.299, p =.001, η_p^2 = .29, where participants took longer to respond to incongruent (M = 774.46 ms, SD = 112.84) than to congruent trials (M = 761.68 ms, SD = 112.28). Neither of the interactions were significant (all p_s>.1).

4.4.2. fMRI

Sustained activity.

No pattern of sustained activations survived statistical corrections, neither all blocks collapsed (vs. rest) or any between-block comparisons.

Transient activity.

First, we ran a full factorial ANOVA including the three block cues, to investigate whether control mechanisms acted differently during social and non-social blocks (see Table 1, Figure 3). To examine the directionality of the results, we extracted the beta estimates for each condition and timebin. Here, we observed a triple interaction of Dimension x Interference x Timebin, where the pattern of the BOLD signal for congruent trials in the identity task contrasted with the other conditions at different time points in

supplementary motor area/middle cingulate cortex (preSMA/SMA/MCC), middle frontal gyrus (MFG) and occipital areas extending to the lingual (LG) and fusiform gyrus (FG). Moreover, we observed a main effect of Dimension, with increased activation at the right middle temporal gyrus (MTG) for emotion and colour tasks, compared to the identity task.

Table 1. Transient activation results for the three dimensions ANOVA.

Label	ANOVA term	Direction	Peak coordinate	F value	*k*
R MTG	Main effect	C/E > I	54, -49, -7	4.19	103
B LG/FG/MOC	Interaction		21, -94, -7	25.11	8103
B preSMA/SMA/ MCC	Interaction		0, -4, 59	10.61	683
L MFG	Interaction		-33, 38, 25	3.52	90

Note: The term Interaction refers to the triple interaction Dimension x Interference x Time. C = Colour trials; E = Emotion trials; I = Identity trials; MTG = Middle Temporal Gyrus; LG = Lingual Gyrus; FG = Fusiform Gyrus; MOC = Middle Occipital Cortex; SMA = Supplementary Motor Area; MCC = Middle Cingulate Gyrus; MFG = Middle Frontal Gyrus.

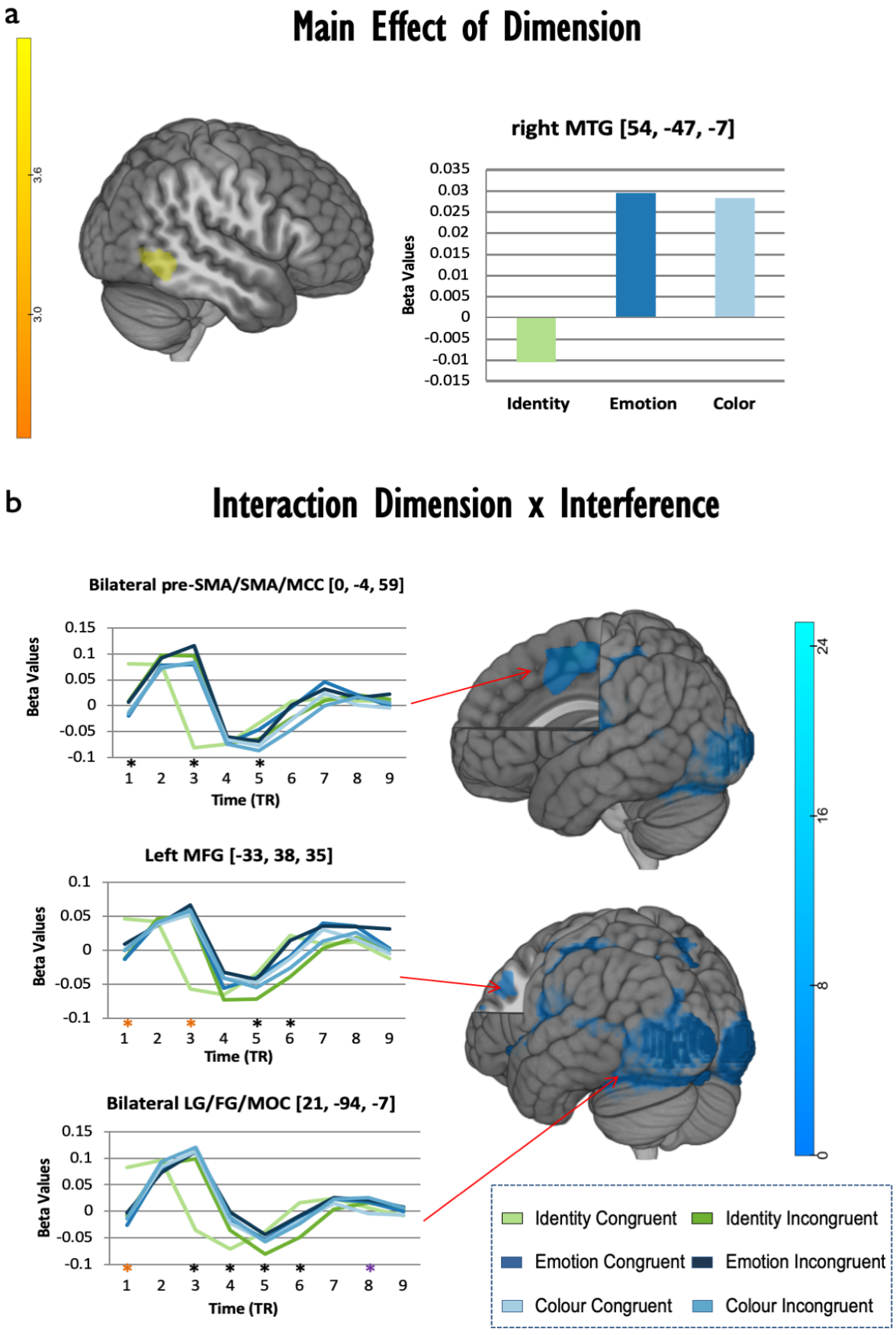
a
Main Effect of Dimension
3.6
3.0
right MTG [54, -47, -7]
Beta Values
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
Identity
Emotion
Color
b
Interaction Dimension x Interference
Bilateral pre-SMA/SMA/MCC [0, -4, 59]
Beta Values
0.15
0.1
0.05
0
-0.05
-0.1
1 2 3 4 5 6 7 8 9
Time (TR)
Left MFG [-33, 38, 35]
Beta Values
0.1
0.05
0
-0.05
-0.1
1 2 3 4 5 6 7 8 9
Time (TR)
Bilateral LG/FG/MOC [21, -94, -7]
Beta Values
0.15
0.1
0.05
0
-0.05
-0.1
1 2 3 4 5 6 7 8 9
Time (TR)
24
16
8
0
Identity Congruent
Identity Incongruent
Emotion Congruent
Emotion Incongruent
Colour Congruent
Colour Incongruent

Figure 3. Results from the three dimensions ANOVA. a) Orange cluster shows the region where the main effect of Dimension was significant. Bars show the hemodynamic response (beta values) for identity (green), emotion (light blue) and colour (blue) trials. b) The blue clusters show the regions where the interaction Dimension x Interference x Time was significant. Lines depict the beta values for congruent and incongruent trials during identity, emotion and colour tasks. Asterisks reflect the timebins where the effect of interference was significant (p<.05) for identity (black), identity and emotion (orange) or colour (purple). Scales reflect peaks of significant t-values (p<.05, FWE-corrected for multiple comparisons).

Results of the first ANOVA suggest that the main differences in the BOLD signal were triggered mainly for the identity task in comparison with the other two. Since one of our main goals relied on studying the different mechanisms underlying the use of social information (personal identity and emotion) to guide trust decisions, we performed a second ANOVA that allows to compare directly the social tasks and the specific interference between these two dimensions. In this social dimensions' ANOVA (see Table 2 and Figure 4), we observed a main effect of Dimension, where emotion trials yielded stronger activation in the inferior frontal gyrus (IFG), medial frontal cortex including superior frontal gyrus and anterior cingulate cortex (extending to MCC), inferior temporal gyrus (ITG), LG, precuneus and middle temporal gyrus (MTG). Moreover, the Interference between both social dimensions engaged lateral PFC (MFG/SFG), SMA, parietal cortex and the precuneus. Last, data showed an interaction between Dimension and Interference, where activation in inferior and superior parietal lobes (IPL/SPL), as well as precentral gyrus and the precuneus decreased for congruence in identity trials, showing the opposite pattern in the emotion task (see Figure 3).

Table 2. Transient activation results for the social dimensions ANOVA.

Label	ANOVA term	Direction	Peak coordinate	F value	*k*
R IFG	Main effect	E > I	54, 35, -1	31.9	304
B mSFG/ACC	Main effect	E > I	0, 47, 35	28.96	798
R ITG/ Cerebellum Crus	Main effect	E > I	54, -58, -13	26.64	397
R PreCuneus/IPL	Main effect	E > I	18, -55, 44	22.91	150
R LG	Main effect	E > I	15, -82, -10	18.82	144
L MTG	Main effect	E > I	-54, -73, 8	17.01	68
R MFG/SFG	Main effect	In > Con	36, 53, 26	27.06	139
L SPL/PostCen	Main effect	In > Con	-15, -37, 53	22.72	121
L PreCuneus	Main effect	In > Con	-9, -73, 35	19.41	75
B preSMA/SMA	Main effect	In > Con	0, 14, 44	18.39	127
L PreCen	Interaction	CE>InE InId>CId	-39, -1, 32	32.18	121
L IPL/SPL, MOC	Interaction	CE>InE InId>CId	-27, -58, 38	26.12	315
R AG/IPL	Interaction	CE>InE InId>CId	27, -55, 44	17.32	71

Note: The term Interaction refers to the interaction Dimension x Interference. E = Emotion trials; I = Identity trials; In = Incongruent trials; Congruent trials; CE = Congruent trials in emotion task; InE = Incongruent trials in emotion task; CId = Congruent trials in identity task; InId = Incongruent trials in identity task; IFG = Inferior Frontal Gyrus; mSFG = Medial Superior Frontal Gyrus; MCC = Middle Cingulate Gyrus; ITG = Inferior Temporal Gyrus; IPL = Inferior Parietal Lobe; LG = Lingual Gyrus; MTG = Middle Temporal Gyrus; AG = Angular Gyrus; MFG = Middle Frontal Gyrus; SPL = Superior Parietal Lobe; PostCen = Post central gyrus; SMA = Supplementary Motor Area; PreCen = Precentral gyrus; MOC = Middle Occipital Cortex.

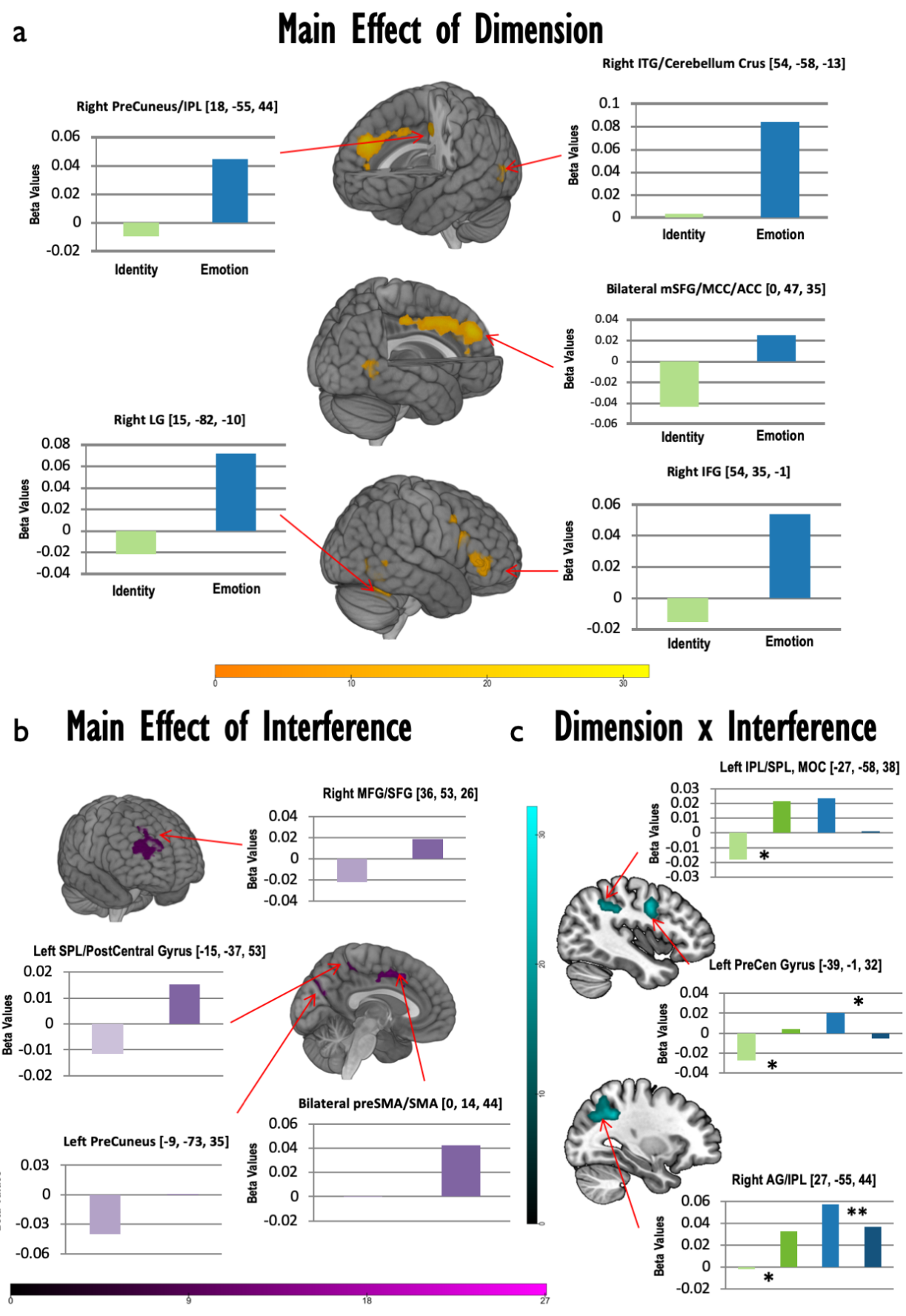
a
Main Effect of Dimension
Right PreCuneus/IPL [18, -55, 44]
Beta Values
0.06
0.04
0.02
0
-0.02
Identity
Emotion
Right ITG/Cerebellum Crus [54, -58, -13]
0.1
0.08
0.06
0.04
0.02
0
Bilateral mSFG/MCC/ACC [0, 47, 35]
0.04
0.02
0
-0.02
-0.04
-0.06
Right LG [15, -82, -10]
0.08
0.06
0.04
0.02
0
-0.02
-0.04
Right IFG [54, 35, -1]
0
10
20
30
b
Main Effect of Interference
Right MFG/SFG [36, 53, 26]
Left SPL/PostCentral Gyrus [-15, -37, 53]
0.02
0.01
-0.01
-0.02
Bilateral preSMA/SMA [0, 14, 44]
Left PreCuneus [-9, -73, 35]
0.03
-0.03
-0.06
0
9
18
27
c
Dimension x Interference
Left IPL/SPL, MOC [-27, -58, 38]
-0.03
*
Left PreCen Gyrus [-39, -1, 32]
**
Right AG/IPL [27, -55, 44]

Figure 4. Results from the social dimensions ANOVA. (a) Orange clusters show the regions where the main effect of Dimension was significant. (b) Violet clusters show the regions where the main effect of Interference was significant. (c) Cyan clusters show the regions where the interaction Dimension x Interference was significant. Bars show the beta values for congruent (light violet) and incongruent (dark violet) trials; congruent (light green) and incongruent (dark green) trials during identity task; congruent (blue) and incongruent (dark blue) trials during emotion task. Asterisks reflect a significant effect of interference in identity and emotion tasks (*p<.05, **p = .053). Scales reflect peaks of significant t-values (p<.05, FWE-corrected for multiple comparisons).

4.5. Discussion

In the present study, we aimed to examine sustained and transient control mechanisms underlying the interference between social and non-social information during interpersonal decisions. Our results showed conflict effects in behaviour and brain activation. We observed transient activation in frontoparietal areas in the presence of interference between social dimensions, and part of the cinguloopercular network, the dACC/mSFG, that was sensitive to the type of social task.

Behavioural data showed general interference effects, where participants were faster and more accurate when the relevant dimension was congruent with the other two. These results extend previous work that found interference markers in non-emotional and emotional conflict (Egner & Hirsch, 2005; Torres-Quesada et al., 2014) and during social decisions when identity and emotion were incongruent (Alguacil et al., 2015), highlighting that both social and non-social information are susceptible to interference from irrelevant dimensions during interpersonal decisions.

At the brain level, we observed differences between identity and the other two dimensions. First, we observed higher activation for emotion and colour tasks, compared to identity in MTG, a region linked to semantic (e.g. González-García et al., 2016; Jefferies, 2013) and social knowledge (Contreras, Banaji, & Mitchell, 2012). On top of that, a triple interaction showed interference effects in identity trials at some time points, where the time course for all conditions was similar except for congruent trials in the identity task. This interaction was manifested in face-processing regions such as the LG and FG (Haxby et al., 2002; Kanwisher & Yovel, 2006) and control-related regions. For instance, the pre-SMA and MCC have been associated with task-rule activation (Crone et al., 2006; De Baene & Brass, 2014) and phasic selective attention (Dosenbach et al., 2007), respectively. Lateral PFC, on the other hand, contributes to conflict resolution amplifying the processing of task relevant information (Egner & Hirsch, 2005). Altogether, this pattern of results could suggest that the identity task required fewer resources than the other two, especially during congruent trials, where the activation decreased in areas associated with emotional conflict and general control-related regions.

In this analysis we could not tease apart interference effects from each specific irrelevant dimension. Since we did not find differences between social and non-social tasks and we were mainly interested in contrasting different sources of social information, we performed a second ANOVA to examine in more detail the specific interactions between only the social dimensions. First of all, we observed increased activation for the emotion

task compared to the identity one, in line with the previous analysis. This time we found task differences in frontoparietal regions associated to cognitive control. Importantly, one of these regions was a cluster in dACC/mSFG, one of the main nodes of the cingulo-opercular network proposed by Dosenbach et al. (2008). These authors associated this region with a role in sustained maintenance of goal-directed tasks. However, our results are in line with previous work (Gratton et al., 2017) that evidenced that cingulo-opercular regions showed transient responses during decision-making. Likewise, we extend this finding to interpersonal decisions (but see also Alguacil et al., *unpublished*). In addition, we observed activation in the precuneus/IPL, in charge of phasic control (Dosenbach et al., 2008) and related to mentalizing processes (Saxe, 2006; van Overwalle, 2006). Similarly, the right IFG is associated with both control (Gratton et al., 2017; Szameitat et al., 2002) and the response to emotional expressions (Thye, Murdaugh, & Kana, 2018). Moreover, the emotion dimension yielded increased activation in the LG, which participates in the contextual integration of faces (Freeman et al., 2015), and in stimuli discrimination whose spatial configuration is associated with more than one possible responses (Sulpizio, Committeri, Lambrey, Berthoz, & Galati, 2013). Overall, these results suggest that the emotion task recruits control mechanisms to maintain task-relevant information, enhancing the processing of facial expressions in specialized areas.

Importantly, this second analysis with the social dimensions allowed us to observe neural markers of interference in regions belonging to the FP control network. Notably, Ruz & Tudela (2011) found a similar cluster in the

SMA/MCC that was coupled with the dACC during conflict from emotional expressions. Our results extend previous work about their role to conflict resolution during interpersonal decisions, where two important sources of social information trigger opposite decisions. Conversely, congruent trials were associated with a deactivation profile in the precuneus. Previous studies have observed coupling between the dACC and the precuneus when emotion was congruent with identity (Alguacil et al., *unpublished)* or was followed by its natural consequences (Ruz & Tudela, 2011). These previous studies explained their results in terms of participants relying on mentalizing processes or processing the identity attributes without emotional interference. Our pattern of results does not exclude this possibility, since previous studies have proposed that deactivated regions could have an important role during task performance (Landsiedel & Gilbert, 2015; Spreng, 2012). An alternative possibility is that considering the role of the precuneus in phasic control (Dosenbach et al., 2008), its deactivation profile could point the lack of conflict between the dimensions.

Notably, the interference between social dimensions interacted with the task in another set of FP regions that represents task-relevant information (e.g. Loose et al., 2017; Woolgar et al., 2016). This pattern is in line with the findings from the first ANOVA, where congruent trials in the identity task show decrease activation in control-related areas. Altogether, we show different activation effects for identity compared to the other dimensions. However, we did not observe differences between the emotion vs. the other two or between social and non-social information. These findings are somewhat intriguing and require further study to determine if different

neural mechanisms underlie the representation of these tasks, or whether the patterns that we observed are due to other variables, such as task difficulty.

Moreover, another goal of this study was to examine the stable maintenance of task-relevant information. However, despite our results in phasic activity, we did not observe any surviving sustained block-related activation. Although this could be due to low statistical power, we followed the recommendation of Petersen & Dubis (2012), who suggested a minimum sample size of 25-30 participants for mixed designs. Other studies have also shown a lack of significant effect for sustained block activation in hybrid designs. In this line, Palenciano et al. (2019b) carried out an instructions following study where they observed sustained activation for novel instructions blocks, but not for practiced ones. This could indicate that the maintenance of control mechanisms is related to difficulty, and our task, even if it was set in a complex decision-making scenario, entailed choices that were rather easy to perform. Moreover, Dubis et al. (2016) observed that sustained cingulo-opercular activation was not present in tasks driven by perceptual information. This could explain the lack of sustained results in our study, where participants relied on dimensions (social: identity/emotion or non-social: colour) that were obtained from perceptual attributes.

Our study is not exempt from some limitations. First, our main goal was to examine control mechanisms in social scenarios. However, the characteristics of the scanner environment make somewhat difficult to engage participants in a real social exchange rather than a mere computer game. Moreover, participants could have not believed the interpersonal setting of the

experiment, and for them all dimensions could be equally relevant of the decision-making. This would explain the lack of differences between social and non-social dimensions in both behaviour and brain data. Also, we used only four identities in the game, which could have made the task rather easy to perform. Perhaps if more partners were included and the relevant social dimensions varied on a trial-by-trial basis, it would increase the task difficulty and we would observe the participation of sustained mechanisms in a more effortful scenario. Another aspect of our study is that our cues were completely predictive, however, our decisions in the real world are made under uncertainty, which has been showed to modulate the impact of social information on decisions in previous interpersonal decision studies (e.g. Ruz et al., 2011).

With all, our puzzling results serve as an incentive to better characterize the neural representation of different types of information during social decisions. In this sense, recent approaches like Multivoxel Pattern Analysis (MVPA; Haxby et al., 2014; Haynes, 2015) and Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) would help us to describe at a fine-grained level how stimuli are organized depending on task-relevant dimension, and how the representation of social knowledge modulates behaviour. Further, in this study we have used a trust game with facial stimuli as the relevant information to guide decisions. Still, we do not know whether these findings can be generalized to other interpersonal scenarios and with a different type of social information. Thus, employing different interpersonal settings and sources of social knowledge (e.g. traits' descriptions) will allow

us to examine to what extent our present result can be generalized to other social scenarios.

CHAPTER 5

EXPERIMENT III

The content of this chapter is in preparation as Díaz-Gutiérrez, P.; Arco, J.E.; Alguacil. S.; González-García, C. & Ruz, M. Neural representation of social expectations during interpersonal decisions. Preprint available at bioRrxiv: https://doi.org/10.1101/355115

5.1. Abstract

Several studies highlight the relevance of prior personal information during social interactions. Such knowledge aids in the prediction of others, and it affects choices even when it is unrelated to their actual behaviour. In this investigation, we aimed to study the neural representation of positive and negative personal expectations, how these impact subsequent choices, and the effect of mismatches between expectations and encountered behaviour. We employed functional Magnetic Resonance in combination with a version of the Ultimatum Game where participants were provided with information about their partners' moral traits previous to their fair or unfair offers. Univariate and multivariate analyses revealed the implication of the supplementary motor area (SMA) and inferior frontal gyrus (IFG) in the representation of expectations about the partners in the game. Further, these regions also represented the valence of expectations, together with the medial prefrontal cortex (mPFC). Importantly, the performance of the classifier in these clusters correlated with the behavioural choice bias to accept more offers following positive descriptions, highlighting the impact of the valence of the expectations on participants' decisions. Altogether, results suggest that the expectations based on social information guide future interpersonal decisions and that the neural representation of such expectations is related to their influence on behaviour.

5.2. Introduction

Decision-making is a crucial constituent of our daily life. A great part of our decisions involves social contexts, where we constantly engage in interactions with others. To make choices that best fit our goals, we weight different sources of information. Within the framework of prediction coding (Friston, 2005), optimal decision-making combines sensory input (*evidence*) with predictions (*priors;* Schwarz et al., 2016; Summerfield and De Lange, 2014). The role of these expectations has been thoroughly examined in perceptual decisions, where several studies have shown pre-activation of target-related brain areas during the expectation period, prior to target onset (e.g., Esterman and Yantis, 2010; González-García et al., 2016; Puri et al., 2009). Also, when making decisions in more complex scenarios (Lopez-Persem et al., 2016), people tend to choose more often and faster the preferred option even when the value of the different alternatives is similar. This leads to suboptimal decisions that do not properly consider potential future outcomes (Fleming, Thomas, & Dolan, 2010). Crucially, this is also the case for interpersonal decisions, which can be biased by several sources of information at different stages of processing (Díaz-Gutiérrez et al., 2017). More recently, there have been proposals linking predictive coding and the representation of social traits in relation to social expectations (e.g., Tamir and Thornton, 2018). Nonetheless, to this date, how prior social information influencing subsequent decision-making is represented in the brain is not well understood. In the current investigation, we aimed to study the neural representation of social expectations during a modified Ultimatum Game (UG; Güth et al., 1982; Moser et al., 2014), where participants receive monetary offers from game partners and decide whether to accept them or not. Acceptance leads to both parts earning their split; whereas no gains are

earned after a rejection. Here, “rational” decisions from an economic point of view should be of acceptance, since you can only earn money. However, choices are strongly influenced by the fairness of the offer (how balanced both halves of the split are). People often show high rejection rates towards unfair offers (Sanfey et al., 2003), which has been explained in terms of inequity-aversion tendencies (Fehr & Camerer, 2007) and punishment (Brañas-Garza, Espín, Exadaktylos, & Herrmann, 2014). Others have emphasized the importance of social norms, and how these impact the perception of fairness (Chang & Sanfey, 2013). In these scenarios, the mechanisms underlying the processing of offers depending on their fairness and participants’ subsequent responses have been extensively studied (for a meta-analysis, see Gabay et al., 2014).

Still, how the brain represents socially relevant priors in interpersonal games is largely unknown. Several studies have described a set of regions (social cognition network) underlying the representation of knowledge that guides social predictions in a broad context (Frith & Frith, 2008), including personal traits, stereotyping, semantic knowledge about people or inferences about others and their mental states (Tamir & Thornton, 2018; Tamir, Thornton, Contreras, & Mitchell, 2016). This network includes the temporoparietal junction (TPJ), superior temporal sulcus (STS), precuneus (PC), anterior temporal lobes (ATL), amygdala and the medial prefrontal cortex (mPFC; Contreras et al., 2013; Frith, 2007; Frith and Frith, 2001; Mitchell et al., 2008). These regions underlie processes such as the understanding of other people's intentions, known as Theory of Mind (ToM; Saxe and Kanwisher, 2003). Similarly, in the context of decisions in social contexts, the mPFC has been related to expectations about others’ behaviour (Corradi-Dell’Acqua, Turri,

Kaufmann, Clément, & Schwartz, 2015). Importantly, prior expectations during social decisions also influence behaviour when they are not followed by their usual consequences. In this line, different studies (Fouragnan et al., 2013; Ruz and Tudela, 2011) have observed increased activation in brain areas associated with cognitive control, such as the ACC and the anterior insula (aI) when expectations about partners do not match their subsequent behaviour. Similarly, Chang and Sanfey (2013) found a relationship between the deviation of the expectations and increased activation in the aI, ACC and SMA. Specifically, in the UG, an increase of activation in the dorsolateral PFC (dlPFC) and aI has been related to participants' reaction to unfair offers (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Sanfey et al., 2003), which has also been interpreted as a violation of what we expect from others.

The personal traits of others are essential components of social representations (Tamir & Thornton, 2018). Their processing has a dynamic nature (Freeman & Ambady, 2011; Stolier et al., 2018) where priors routed in stereotypes modify and interact with perceptual processes (Stolier & Freeman, 2016, 2017). Further, these personality traits can be decomposed in three different dimensions: rationality, social impact and, crucially to our investigation, valence (positive vs. negative; Tamir and Thornton, 2018; Thornton and Mitchell, 2017). The representation of the character of others in association with positive or negative information has been shown to be an important source of bias in interpersonal decisions (Díaz-Gutiérrez et al., 2017). For instance, Delgado et al. (2005), found that participants trusted partners associated with positive moral traits more than those having negative ones. Furthermore, a variety of studies employing the UG paradigm have observed that participants tend to accept

more offers from partners associated with positive descriptions, compared to negative ones (Gaertig et al., 2012). This tendency is steeper when participants navigate uncertain scenarios (Ruz et al., 2011). Moreover, in this context, the use of high-density electroencephalography (EEG) has shown that negative descriptions of partners lead to a higher amplitude of the medial frontal negativity (MFN; associated with the evaluation of outcomes, Hajcak et al., 2006; Yeung and Sanfey, 2004) when decisions are made (Moser et al., 2014). This data indicates how, regardless of fairness, people evaluate offers as more negative when they come from a disagreeable partner. Such knowledge about personal traits has been suggested to be integrated by the mPFC (Van Overwalle, 2009). This area increases its coupling with other regions responding to specific traits (Hassabis et al., 2014), and shows heightened activation when a partner's behaviour violates previous trait implications (Ma et al., 2012).

Despite the key relevance of valence in psychological theories and its marked impact on social decision-making, its representation at the neural level and its effect on subsequent choices are not well understood (Barrett & Bliss-Moreau, 2009). Results of a recent meta-analysis (Lindquist, Satpute, Wager, Weber, & Barrett, 2015) provide evidence of a general recruitment of a set of regions for valenced versus neutral information, including the bilateral aI, the ventral and dorsal portions of the medial PFC (vm/dmPFC), the dorsal ACC, SMA, and lateral PFC, which are associated with the "salience network" (Menon & Uddin, 2010; Seeley et al., 2007) and some of them also related to cognitive control (Brass & von Cramon, 2004; Dosenbach et al., 2008). Lindquist et al., (2015) found that the vmPFC/ACC was more frequently activated in positive vs.

negative than in positive vs. neutral contrasts, which could indicate that these regions represent valence information along a single bipolar dimension. However, multivariate analysis (MVPA; Haxby et al., 2014) did not indicate distinctive patterns sensitive to valence discrimination (positive vs. neutral or negative vs. neutral), which suggests that affect is mainly represented flexibly in a valence-general set of regions (Lindquist et al., 2015).

In the current functional Magnetic Resonance Imaging (fMRI) study, we employed a modified version of the UG (Gaertig et al., 2012) to investigate how socially relevant priors represented by the valence of personal descriptions of the partners, bias interpersonal economic choices. Specifically, we aimed to study which neural regions code for the generation and maintenance of positive and negative expectations about other people. Furthermore, we also wanted to assess how these expectations bias decisions. We expected to find specific neural representations underlying the expectations about the partners, with different patterns depending on the valence of these predictions (Lindquist et al., 2015). Specifically, we hypothesized that these patterns would be represented in regions related to social cognition and priors in decision-making (Contreras et al., 2012; González-García et al., 2016; Saxe & Kanwisher, 2003). Last, we intended to ascertain which neural mechanisms were engaged when there is a mismatch between personal expectations and the partners' behaviour. We predicted that control-related areas would be engaged when the valenced description was not congruent with the subsequent partner's behaviour.

5.3. Methods

5.3.1. Participants

Twenty-four volunteers were recruited from the University of Granada (M = 21.08, SD = 2.92, 12 men), matching a sample size previously employed in previous studies with the same paradigm (Moser et al. 2014), and similar to other fMRI studies employing the UG (Chang and Sanfey, 2013; Grecucci, Giorgetta, Bonini & Sanfey, 2013). All participants were right-handed with normal or corrected vision and received economic remuneration (20-25 Euros, proportionally to their acceptance rates). Participants signed a consent form approved by the Ethics Committee of the University of Granada.

5.3.2. Apparatus and stimuli

We employed 16 adjectives used in previous studies (Gaertig et al., 2012; Moser et al., 2014; Ruz et al., 2011; see Table 1) as trait-valenced descriptions of the game proposers, extracted from the Spanish translation of the Affective Norms for English Words database (ANEW; Redondo et al., 2007). Half of the adjectives were positive (M= 7.65 valence, SD= 0.43), and the other half were negative (M= 2.3 valence, SD= 0.67). All words were matched in arousal (M= 5.69, SD= 0.76), number of letters (M= 6.19, SD= 1.42) and frequency of use (M= 20.19, SD= 18.47). In addition, we employed numbers from 1 to 9 (two in each trial) in black colour to represent different monetary offers. Stimuli were controlled and presented by E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). Inside the scanner, the task was projected on a screen visible to participants through a set of mirrors placed on the radiofrequency coil.

5.3.3. Task and procedure

To add credibility to the interpersonal game setting, participants were told that they were about to receive offers made by real participants in a study of a previous collaboration with a foreign university. Furthermore, to engage participants in the game as a real social scenario, prior to the scanner they performed two tasks in which they had to make economic offers that would be used for prospective participants. In one of the tasks, participants acted as proposers, filling a questionnaire where they had to make offers for 16 different unknown partners, who would be involved in future experimental games. Here, they had to split 10 Euros into two parts, one for themselves and the other for their partners. Additionally, in a second task, they played a short version of the Dictator Game (Kahneman, Knetsch, & Thaler, 1986), where they decided how to divide another 10 Euros between themselves and an anonymous partner, who would have a merely passive role concerning the output of the offer. Moreover, participants were told that the offers that they were about to see in the scanner were each provided by a different partner who previously performed the same tasks as they did before the scanner, and therefore, the offers were real examples of others participants' responses when acting as proposers. Participants were informed that each offer would be preceded by a word that had been obtained as an output from a series of personality and social questionnaires filled by their partners. These adjectives described their partners in some way (see Table 1). Choices made by participants had an influence in their final payment, as they were informed that it varied (20-25 Euros) according to their choices during the game in the scanner. In a post-scanning debriefing session, none of the participants reported suspicions regarding the

background story of this procedure, which has also been used successfully in other settings (e.g. Correa et al., 2017).

Table 1. List of adjectives employed in the task.

Positive words	Negative words
Friend	Criminal
Generous	Cruel
Honest	Disloyal
Honourable	False
Humble	Guilty
Kind	Hostile
Loyal	Selfish
Warm	Traitor

In the scanner, participants played the role of the responder in a modified UG (e.g., Gaertig et al., 2012), deciding whether to accept or reject monetary offers made by different partners (proposers). If they accepted the offer, both parts earned their respective splits, whereas if they rejected it, neither of them earned money from that exchange. Offers consisted of splits of 10 Euros, which could be fair (5/5, 4/6) or unfair (3/7, 2/8, 1/9). The number presented at the left on the screen was always the amount of money given to the participant, and the one on the right side was the one proposed by the partners for themselves.

Personal information about the partners was included as adjectives with different valence. A third of these descriptions were positive, another third negative, and the last third was neutral, represented by text indicating the

absence of information about that partner ("no test"). The valence of the adjectives was unrelated (50%) to the fair-unfair nature of the offer. The order of the offers and adjectives was randomized, and each type of personal information (positive, negative, no information) preceded each offer equally within and across runs. Decision-response associations were counterbalanced between participants.

Participants performed a total of 192 trials, arranged in 8 runs (24 trials per run). In each run, a start cue of 6 s was followed by 24 trials. Each trial (see Figure 1) started with an adjective for 1 s (mean = 2.98°), preceding a jittered interval lasting 5.5 s on average (4-7 s, +/0.76°). Then, the offer appeared for 0.5 s (1.87°), followed by a second jittered interval (mean = 5.5 s; 4-7 s, +/0.76°). Overall, the task lasted 41 minutes approximately.

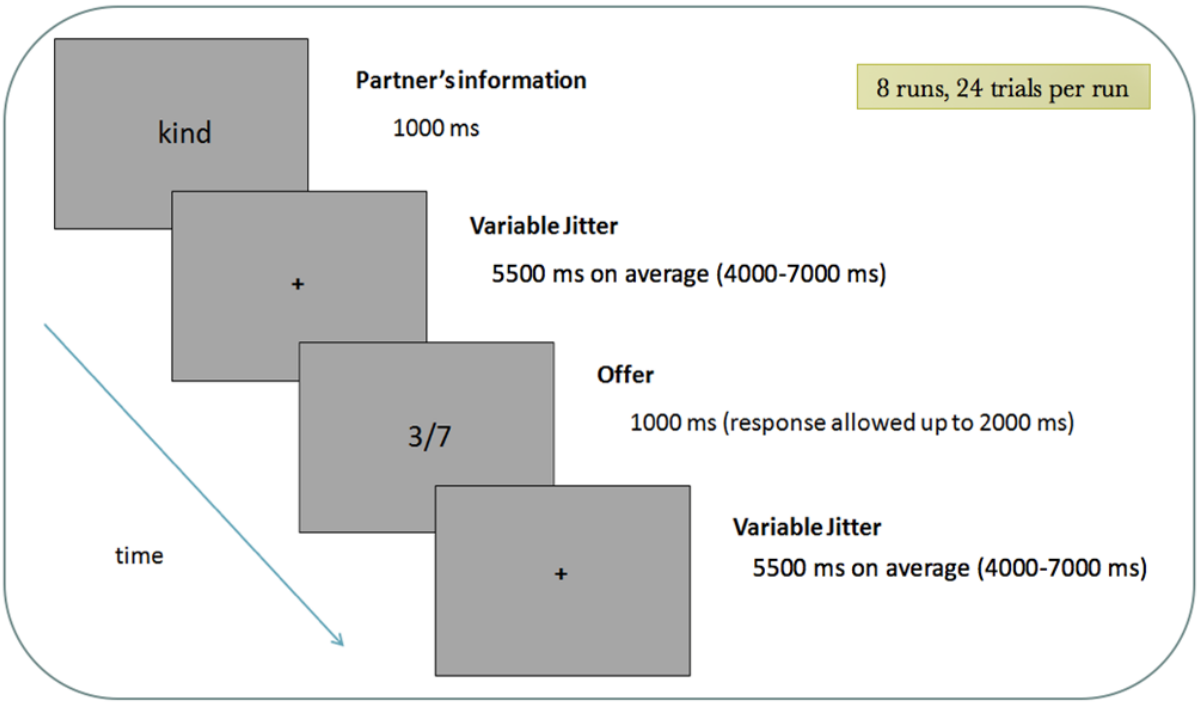

Figure 1. Sequence of events in a trial. The task varied the Valence of the partner's information (Positive, Negative, No information) and the Fairness of the offer (Fair/Unfair).

5.3.4. Image acquisition and preprocessing

MRI images were acquired using a Siemens Magnetom TrioTim 3T scanner, located at the Mind, Brain and Behaviour Research Centre in Granada. Functional images were obtained with a T2*-weighted echo planar imaging (EPI) sequence, with a TR of 2000 ms. Thirty-two descendent slices with a thickness of 3.5 mm (20% gap) were extracted (TE = 30 ms, flip angle = 80 º, voxel size of 3.5 mm^3). The sequence was divided into 8 runs, consisting of 166 volumes each. After the functional sessions, a structural image of each participant with a high-resolution T1-weighted sequence (TR = 1900 ms; TE = 2.52 ms; flip angle = 9º, voxel size of 1 mm^3) was acquired.

Data were preprocessed with SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The first three volumes of each run were discarded to allow the signal to stabilize. Images were realigned and unwarped to correct for head motion, followed by slice-timing correction. Afterwards, T1 images were coregistered with the realigned functional images. Then, functional images were spatially normalized according to the standard Montreal Neurological Institute (MNI) template and smoothed employing an 8 mm Gaussian kernel. Low-frequency artefacts were removed using a 128 high-pass filter. On the other hand, data for multivariate analyses was only head-motion and slice-time corrected and coregistered.

5.3.5. Univariate analyses

First-level analyses were conducted for each participant, following a General Linear Model in SPM12. We employed an event-related design, where activity was modelled using regressors for each valence type of adjective and for the

offers. The estimated model included three regressors for the Words (positive, negative, no information) and six for the Offers (Fair offers_Positive, Fair offers_Negative, Fair offers_Neutral, Unfair offers_Positive, Unfair offers_Negative, Unfair offers_Neutral). Note that since decisions were made when the offers appeared, and that responses (choices) showed a strong dependency on offer fairness, offer fairness and decisions cannot be modelled separately. Given our research questions, we modelled the offer events considering their fairness and not participants' choices. Regressors were convolved with a standard hemodynamic response, with adjectives modelled with their duration (1 s + jitter), and offers modelled as events with zero duration. This temporal difference is accounted by the fact that the words describing the partners trigger preparatory processes, which extend in time (Bode and Haynes, 2009; Sakai, 2008). Conversely, this is not the case for the offer, as once the participant chooses, the process ends.

At the second level of analysis, *t*-tests were conducted for comparisons related to the presence of expectations (information about the partner > no information), the valence of the information (positive > negative, negative > positive) and the fairness of the offer (fair > unfair, unfair > fair). We also carried out contrasts for congruence effects between the events, where we had congruent (positive descriptions followed by fair offers, negative descriptions followed by unfair offers) and incongruent trials (positive descriptions followed by unfair offers, negative descriptions followed by fair offers). To control for false positives at the group level, we employed permutations tests with statistical non-parametric mapping (SnPM13, http://warwick.ac.uk/snpm) and 5000 permutations. We performed cluster-wise inference on the resulting

voxels with a cluster-forming threshold of 0.001, which was later used to obtain significant clusters (FWE corrected at $p<0.05$).

5.3.6 Multivariate analyses

We performed MVPA to examine the brain areas representing the valence of the expectations, that is, the regions containing information about whether the partners were described with positive vs. negative adjectives. To this end, we performed a whole brain searchlight (Kriegeskorte et al., 2006) on the realigned images (prior to normalization). We employed The Decoding Toolbox (TDT; Hebart et al., 2015), to create 12-mm radius spheres, where linear support vector machine classifiers (C=1; Pereira et al., 2009) were trained and tested using a leave-one-out cross-validation scheme, employing the data from the 8 scanning runs (training was performed with data from 7 runs and tested in the remaining run, in an iterative fashion). We used a Least-Squares Separate model (LSS; Turner, 2010) to reduce collinearity between regressors (Abdulrahman & Henson, 2016; Arco et al., 2018). This approach fits the standard hemodynamic response to two regressors: one for the current trial (positive/negative adjective) and a second one for all the remaining trials. As in the previous analyses, adjective regressors were modelled with their duration (1 s + jitter) and offers with zero duration. Consequently, the output of this model was one beta image per event (total = 128 images, 64 for each type of adjective, 112 for training and 16 for testing in each iteration). Afterwards, at the group level, non-parametrical statistical analyses were performed on the resulting accuracy maps following the method proposed by Stelzer et al. (2013) for MVPA data. We permuted the labels and trained the classifier 100 times for each subject. The resulting maps were then normalized to an MNI space. Afterwards,

we randomly picked one of these maps per each participant and averaged them, obtaining a map of group accuracies. This procedure was repeated 50000 times, building an empirical chance distribution for each voxel position and selecting the 50th greatest value, which corresponds to the threshold that marks the statistical significance. Only the voxels that surpassed this were considered significant. The resulting map was FWE corrected at 0.05, computing previously the cluster size that matched this value from the clusters obtained in the empirical distribution.

5.3.7. Relationship between decoding accuracy and choices

To examine the extent to which the accuracy of decoding between the two types of adjectives (positive vs. negative) related to the decisions made by participants, we performed a correlation analysis between the individual bias index and mean decoding accuracy values from each significant cluster in the MVPA described above. To obtain this behavioural index, for each participant, we subtracted the average acceptance rate following negative descriptions from the average acceptance rate after positive descriptions (regardless of the nature of the offer). For each subject, we performed a one-tailed (right) Spearman's correlation between the behavioural index and the decoding accuracy from each significant cluster (Bonferroni-corrected for multiple comparisons).

5.4. Results

5.4.1. Behavioural data

Acceptance rates (AR) and reaction times (RTs) were analysed in a Repeated Measures ANOVA, with Offers (fair/unfair) and Valence of the descriptions

(positive, negative, neutral) as factors. The Greenhouse-Geisser correction was applied when sphericity was violated.

5.4.1.1. Acceptance rates

Participants responded on 100% of the trials. Data showed (see Figure 2) a main effect of Offer $F_{1,23} = 74.50$, $p < .001$, $\eta_p^2 = .764$, where fair offers were accepted more often ($M = 84.09\%$; $SD = 22.10$) than unfair ones ($M = 24.18\%$; $SD = 24.10$). Valence was also significant, $F_{2,22} = 13.735$, $p = .001$, $\eta_p^2 = .374$. Participants accepted more offers when they were preceded by a positive description of the partner ($M = 59.39\%$; $SD = 23.09$), than when there was no information ($M = 56.31\%$; $SD = 21.89$) or this was negative ($M = 46.70\%$; $SD = 24.33$). Planned comparisons revealed that these differences were significant between all pairs (all $p_s<.05$). Finally, the Offer X Valence interaction was also significant, $F_{2,22} = 4.262$, $p = .033$, $\eta_p^2 = .156$. Planned comparisons showed that for fair offers, there was no difference between positive and neutral information ($p = .399$), whereas for unfair offers, there was no difference in acceptance rates between negative and neutral information ($p = .074$).

5.4.1.2. Reaction times

Results showed (see Figure 2) a main effect of Offer $F_{1,23} = 22.489$, $p < .001$, $\eta_p^2 = .494$, where participants took longer to respond to unfair ($M = 1023.53$ ms; $SD = 373.10$ ms) than to fair offers ($M = 925.62$ ms; $SD = 309.57$ ms). Neither Valence, $F_{2,22} = 1.05$, $p = .341$, or its interaction with Fairness, $F_{2,22} = 1.956$, $p = .168$ were significant. In addition, to measure the influence of expectations on participant's responses (Ruz et al., 2011), we ran an ANOVA where we included

the valence of the descriptions and the decision (accept, reject) made to the offers. Here, we did not find any effect of Valence, $F<1$, but we found significant effects of Decision, $F_{1,23} = 5.519$, $p = .028$, $\eta_p^2 = .194$, since participants were faster to accept ($M = 951.37$ ms; $SD = 356.01$ ms) than to reject the offers ($M = 988.97$ ms; $SD = 316.91$ ms). Furthermore, data showed an interaction Valence X Decision, $F_{2,22} = 4.23$, $p = .025$, $\eta_p^2 = .155$, replicating previous findings (Gaertig et al., 2012; Ruz et al., 2011). Planned comparisons indicated that these differences in RT for responses took place only after positive, $F_{1,23} = 13.997$, $p = .001$, $\eta_p^2 = .378$ (Accept: $M = 927.60$ ms, $SD = 297.37$ ms; Reject: $M = 993.91$ ms, $SD = 335.52$ ms), and neutral descriptions, $F_{1,23} = 4.504$, $p = .045$, $\eta_p^2 = .165$ (Accept: $M = 955.8$ ms, $SD = 304.96$ ms; Reject: $M = 987.80$ ms, $SD = 328.48$ ms), but not for negative descriptions, $F<1$.

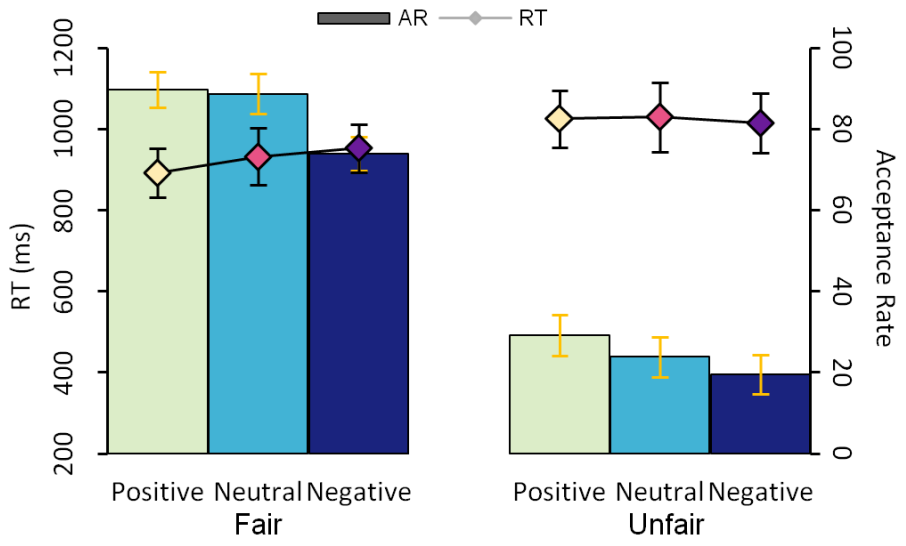

Figure 2. Acceptance Rates (AR, bars) and reaction times (RT, lines) to fair and unfair offers preceded by positive, negative and neutral descriptions of the partner (error bars represent S.E.M).

5.4.2. Neuroimaging data

5.4.2.1. Univariate results

Expectations

During the presentation of the description and the time interval that followed, that is, when participants had personal information to **generate expectations** [(Positive adjective ∩ Negative adjective) > No Information], we observed a cluster of activity (see Figure 3) in the left dorsal aI (k = 109; -33, 21, 4) and bilateral Supplementary Motor Cortex (SMA; k = 138; -8, 11, 53; see Fig. 3). Additionally, the right inferior parietal lobe (right IPL) showed higher activity (k = 264; 55, -35, 53) for **positive descriptions** compared to negative ones. No cluster surpassed the statistical threshold (p>0.05) for the opposite contrast.

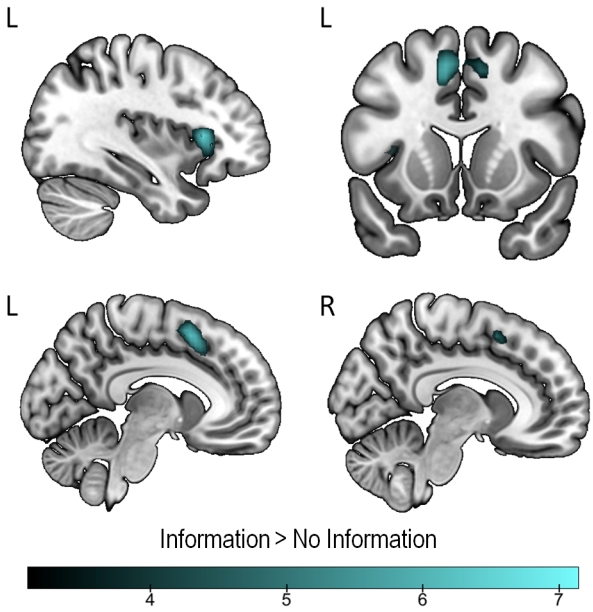

Figure 3. Univariate results during the expectation period. Scales reflect peaks of significant t-values (p<.05, FWE-corrected for multiple comparisons).

During *offer processing*, the previous presentation of **personal information** about the partner [(Offer_Pos ∩ Offer_Neg > Offer_Neu] yielded again

significant activity involving the bilateral dorsal aI and right SMA (k = 23349; -33, 21, 4).

To check whether the regions related to personal information were the same during the presentation of the valenced adjectives and during the presentation of the offer (positive and negative > neutral in both cases), we ran a conjunction analysis with the regions significant in both contrasts (Nichols, Brett, Andersson, Wager, & Poline, 2005). Similar to each contrast individually, we observed two clusters: one in the left IFG/aI (k = 93; -3, 21, 0) and one involving bilateral SMA (k = 126; -5, 18, 53), suggesting that both areas increased their activation during the expectation and offer stages.

Offer fairness

Fair offers (Fair > Unfair) generated activity (see Figure 4) in the right medial frontal gyrus (mFG) and ACC (k = 171; 6, 39, -14), while the opposite contrast (unfair > fair) did not yield any significant clusters (p>0.05). Furthermore, we examined neural responses depending on whether previous expectations were matched or not by the nature (fair vs. unfair) of the offer. Here, **congruence** (see Figure 4) between expectations and offer (Congruent > Neutral) showed a cluster of activity in right cerebellum (right Crus; k = 153; 17, -88, -32). Conversely, **incongruence** (see Figure 4) between expectations and offer (Incongruent > Neutral) yielded activations in the right medial Superior Frontal Gyrus (mSFG) and its lateral portion bilaterally (k = 401; 13, 39, 56), as well as in left IFG (k = 177; -54, 39, 0). Lastly, regarding possible conflict effects, a comparison between **incongruent vs. congruent** trials showed (see Figure 4)

clusters of bilateral activity in the IFG/aI (k = 232; -43, 25, -11/ k = 140; 34, 35, 4).

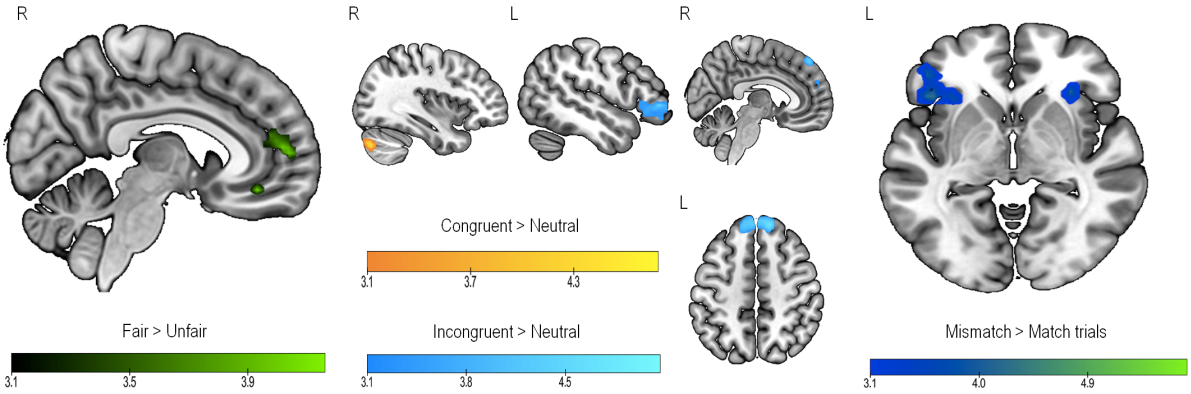

Figure 4. Univariate results for the offer. Scales reflect peaks of significant t-values (p<.05, FWE-corrected for multiple comparisons).

5.4.2.2. Multivariate results

Valence of expectations' classification

Expectations about the partner (positive vs. negative information) showed distinct patterns of neural activity (see Figure 5), in a cluster including the left inferior and middle frontal gyrus (IFG/MFG) and aI (k = 319; -46.5, 28, -32.2), the bilateral medial frontal gyrus and ACC (k = 483; 6, 21, -19.6) and the bilateral middle cingulate cortex (MCC) and SMA (k = 339; -4.5, 14, 35).

Although the same comparisons (positive vs. negative) in univariate GLM only yielded a significant cluster activation in the IPL for positive > negative expectations, we ran a conjunction analysis (Nichols, Brett, Andersson, Wager & Poline, 2005) to test whether the regions that increased their activation during the presentation of the adjectives (positive & negative > neutral) were the same as those that contained relevant information about the valence (as reflected by

multivariate results). For this, we computed the intersection between the group maps from both contrasts. Results showed two clusters (see Figure 5): one in the left IFG/aI (k = 56; -36, 25, 0) and one involving bilateral SMA (k = 69; -8, 18, 46).

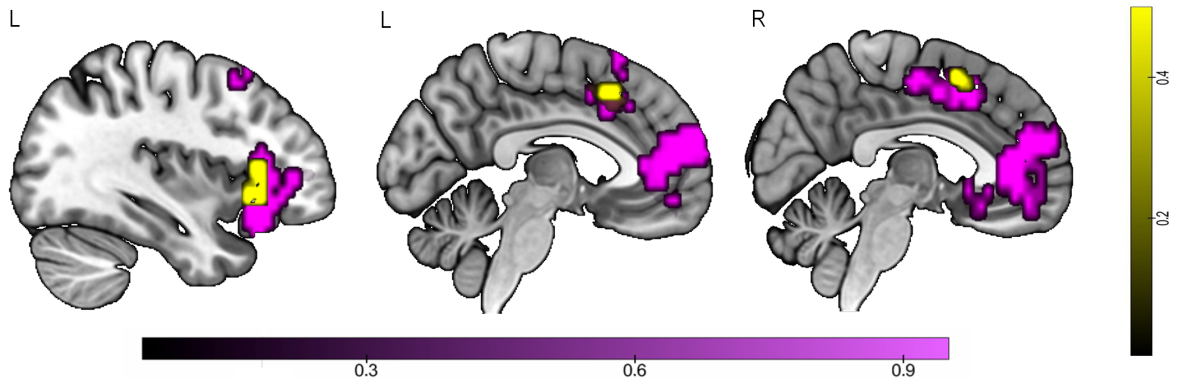

Figure 5. Multivariate results (violet). Different neural patterns for the valence (positive vs. negative) of the adjective during the expectation stage. Scales reflect corrected p-values (<.05). Regions significantly active both during univariate and multivariate analyses are highlighted in yellow.

Importantly, the valence of the description and participants' choices influenced acceptance rates, which were higher for positive descriptions than negative descriptions. This generated potential confounds in the previous decoding. The association between hand and decision (left/right, acceptance/rejection) was fully counterbalanced across participants, but remained constant for each of them. Therefore, the classifier could use response information (accept vs. reject) when decoding valence. To clarify this issue, we performed a response classification at the offer period (following the same procedure as for the valence decoding, see section 5.3.6 Multivariate analysis). Then, we run a conjunction analysis, computing the intersection between valence and response

group maps to examine whether the regions containing relevant information about the valence were the same as those representing participants' decisions (accept vs. reject). Here, we observed that a cluster in bilateral SMA (k = 95; -1, 7, 48) resulted significant for both classification analyses.

Correlation between decoding accuracy and the bias index

To explore the link between behaviour and the decoding results, we correlated the mean decoding accuracies (positive vs. negative) for each significant cluster in the MVPA with the bias index for each participant, which represents how much influence the valence of the adjectives had on their choices. This analysis yielded significant positive correlations between the decoding accuracy for the descriptions' valence and the behavioural bias in all 3 significant clusters (see Figure 6): the left IFG/MFG and aI (r = .42; p = .02), bilateral mFG/ACC (r = .44; p = .015), and the left MCC/SMA (r = .53; p = .0038). Hence, the better the activation patterns in these regions discriminated between the valence of the partners' information, the larger the effect of valenced information on subsequent choices.

Second, to further examine the extent to which participants' responses were related to valence decoding, we ran an additional correlation analysis following the same approach, this time to examine the link between valence' decoding results and the response made by participants (acceptance or rejection of the offer). Therefore, for each participant, we calculated their average acceptance rate in general, regardless of the valence of the expectation and the fairness of the offers. Here, results showed no significant correlation between any of the ROIs mean accuracies and general acceptance rate per participant (all ps>.39),

which supports the specificity of the link between valenced expectations and choices.

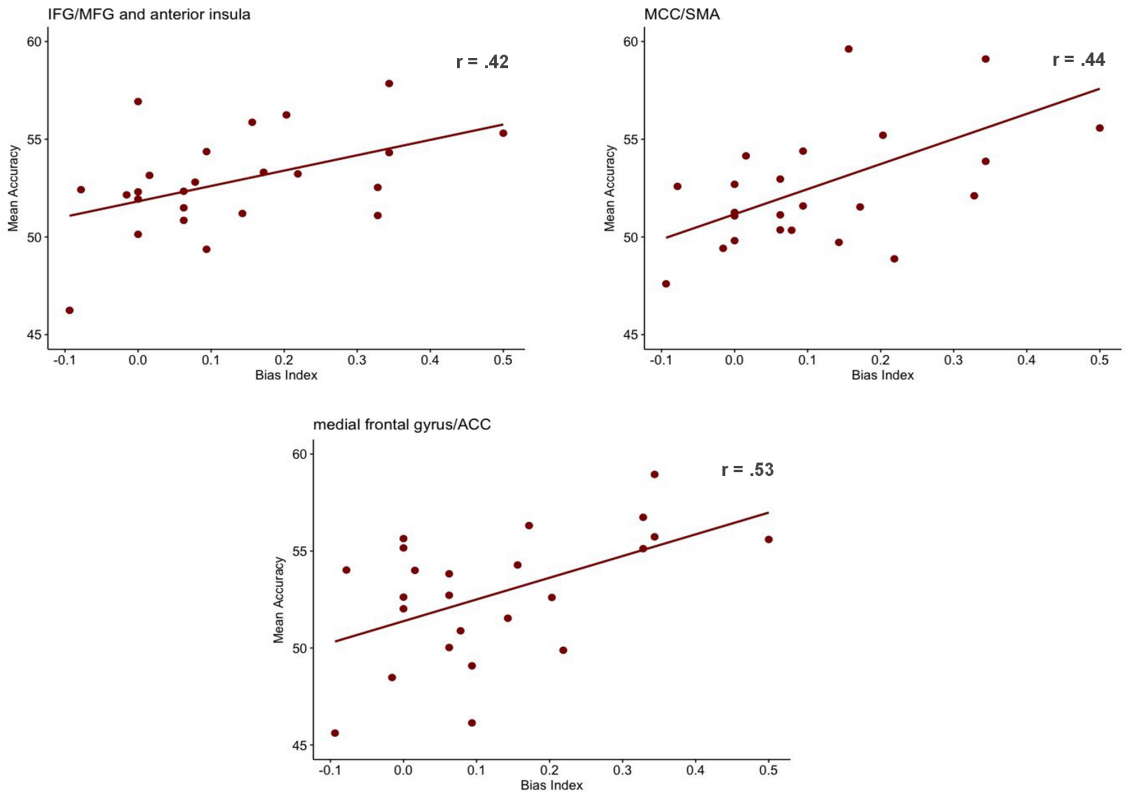

Figure 6. Scatter plots showing significant correlations between mean decoding accuracies in each cluster and the behavioural index. IFG: Inferior frontal gyrus. MFG: Middle frontal gyrus. ACC: Anterior Cingulate Cortex. MCC: Middle Cingulate Cortex. SMA: Supplementary Motor Area.

5.5. Discussion

Our study investigated the neural basis of social valenced expectations during an interpersonal UG. Results revealed how social information about other people bias subsequent economic choices, as well as the brain regions increasing their activity during the maintenance of expectations with a later impact on behaviour. Furthermore, decoding analysis allowed us to observe the areas that

represent the content of such expectations, and how the strength of these social valenced representations influenced ulterior decisions.

The UG employed showed a clear behavioural effect of interpersonal expectations, where positive descriptions of others led participants to higher acceptance rates compared to negative ones. Further, expectations influenced fair and unfair offers differently, where there was no difference between having positive expectations and having no information about the partner for fair offers, and no difference between lack of information and negative information for unfair ones. Additionally, the impact of the expectations was reflected on the speed of choices, where people needed more time to reject offers after positive (or neutral) expectations. This data replicates previous results, where valenced descriptions of partners in bargaining scenarios bias economic choices (Gaertig et al., 2012; Moser et al., 2014; Ruz et al., 2011), emphasizing the role of expectations (Sanfey, 2009) and valenced morality in decision-making (Barrett & Bliss-Moreau, 2009). Results also fit with evidence from non-social decision making (Fleming et al., 2010; Lopez-Persem et al., 2016), highlighting the impact of priors on subsequent decisions. Overall, the behavioural pattern of choices observed supports the utility of the experimental paradigm to induce interpersonal valenced expectations about others that bias subsequent choices made to the same set of objective behaviour (offers made by partners).

Several regions increased their activation when participants held in mind social expectations about game partners. This information engaged the SMA and the dorsal aI, which were also active at the offer stage. These are regions previously

related to preparation processes (Brass & von Cramon, 2004), as well as sustained (Dosenbach et al., 2008) and transient (Menon & Uddin, 2010; Sridharan, Levitin, & Menon, 2008) top-down control, in paradigms where participants use cue-related information to perform tasks of different nature on subsequent targets. In the current context, these areas may be involved in using the interpersonal information contained in the cue to guide or bias the action towards a certain choice, according to the valence of the expectation. This explanation fits with the pattern of behavioural results obtained, where positive expectations increased the acceptance of the offers, as well as speeded up choices to fair offers matching prior expectations. However, univariate contrasts between the words containing positive vs. negative information, in stark contrast with behavioural outcomes, showed effects restricted on a cluster in the IPL for the positive vs. negative contrast. This region has been related to the simulation of others' action in shared representations (Van Overwalle, 2009), and a part of our cluster it is included in the TPJ (e.g., Scholz et al., 2009), which plays a main role in ToM (Saxe & Kanwisher, 2003). The increase of activation in this region for positive expectations could indicate a higher reliance on positive descriptions by the ToM processes involved in our task. This fits with the pattern found in RTs where only positive expectations speeded acceptance choices, whereas negative description did not speed rejections. Further research will be needed to replicate this imbalance of information and to better understand the nature of its underlying brain processes.

Importantly, the use of a multivariate approach, based on classification analysis (i.e., MVPA), allowed us to observe which regions are sensitive to whether the expectations about the partners are positive or negative. This is especially

relevant since previous work has indicated how valence differences at a neural level are particularly hard to observe (Lindquist et al., 2015). These areas included the SMA/MCC, IFG/MPFC and mPFC/ACC. There was no difference in RT between positive and negative conditions (see Behavioural data, section 3.1.), and therefore, we can rule out the possibility that the classifier was mistakenly discriminating faster and slower conditions, and that it correctly differentiated patterns referring to valence instead.

The SMA has been previously associated with general preparatory processes (Brass & von Cramon, 2004), although some studies have also been able to decode specific task sets in this region (Bode & Haynes, 2009; Crittenden, Mitchell, & Duncan, 2015). In social scenarios, Chang and Sanfey (2013) found a relationship between the activity in the SMA and the deviation of previous expectations. This region has also been linked to the unspecific representation of valence (Lindquist et al., 2015). Our conjunction analysis shows that part of the SMA increases its activity during the expectation period and also shows different patterns depending on the valence of the expectation. This data could suggest that the SMA carries a role of general preparation, but it also contains specific fine information relevant to the task. In addition, we observe overlapping activation with the response classification, which suggest that this region also contains some information about participants' responses. The MCC, on the other hand, has been associated with an increase of efficiency in decision-making, being involved in the anticipation and consequent expectations of outcomes in a variety of non-social tasks (Vogt, 2016). Further, it has also been related to the prediction and monitoring of outcomes in social decisions (Apps, Lockwood, & Balsters, 2013), and it may play a similar role in our study.

On the other hand, the patterns of activity in a cluster involving the lateral prefrontal cortex (lPFC), including the IFG and MPFC, also discriminated the valence of the expectations. Interestingly, these areas were part of a large cluster that also increased their activation during the maintenance of social information, as revealed by univariate results. In non-social paradigms, the lateral PFC has been related to working memory maintenance (Morgan, Jackson, Van Koningsbruggen, Shapiro, & Linden, 2013; Sala, Rämä, & Courtney, 2003) and other forms of cognitive control (e.g., Reverberi et al., 2012). The IFG specifically has been also associated with the selection of semantic information (Jefferies, 2013; Wagner, Paré-Blagoev, Clark, & Poldrack, 2001), and it is also involved in the expectation to perform different non-social tasks employing verbal material (e.g., González-García et al., 2017; Sakai and Passingham, 2006). Notably, our results extend this role to a social context (see also Filkowski et al., 2016; Thye et al., 2018; Van Overwalle, 2009), where verbal information is used to generate positive or negative expectations about game partners, by showing that the pattern of activity in this frontal region differs depending on the nature of the information used to predict the proximal behaviour of others.

Interestingly, a region that did not increase its overall activation during the expectation period contained patterns related to the valence of the predictions, the mPFC/ACC. Crucially, this area overlaps with the region isolated in the meta-analysis by Lindquist et al., (2015), where they linked its activity with a bipolar representation of valence. On a broader context, this region is part of the social cognition network, associated with mentalizing processes (Koster-Hale &

Saxe, 2013; Tamir et al., 2016), and behaviour guided by social cues, along with the ACC. Previous studies relate the mPFC with predictions about others' desires (Corradi-Dell'Acqua et al., 2015), and priors during valued decisions (Lopez-Persem et al., 2016). Additionally, Van Overwalle (2009) linked this region to the integration of personal traits, and it has been extensively associated with the representation of intentions as well (Haynes et al., 2007). In our experiment, this area maintains valenced information about others, which is later employed to bias decisions about their economic offers, as corroborated by the correlation with choices discussed in the following paragraph.

The association between a brain region and a given behaviour is strengthened when a link can be observed between the fidelity of a pattern of activity and the behavioural outcome studied (Naselaris, Kay, Nishimoto, & Gallant, 2011; Tong & Pratte, 2012). To find this evidence we obtained, for each participant, a bias index representing how much the valence of the personal information influenced their choices and correlated this index with the accuracy of the classifier in disentangling the patterns generated by positive vs. negative words. We observed a positive correlation between these two factors in the three clusters sensitive to the valence of expectations. Thus, the better the classifier distinguished between descriptions of different valence, the more people tended to accept offers preceded by positive compared to negative descriptions. These results strongly suggest that these valenced representations were used to weight posterior acceptance or rejection decisions to the same set of objective offers, biasing behaviour.

We could also observe the effect of expectations by studying the brain activity generated by offers that matched or mismatched previous expectations, that is, fair and unfair offers preceded by descriptions of the same or opposing valence. Here we found cerebellum activity when fair offers were preceded by positive descriptions and unfair ones followed negative adjectives. This region is associated with prediction in a variety of contexts, such as language (Lesage, Hansen, & Miall, 2017; Pleger & Timmann, 2018) or social cognition (Van Overwalle, Baetens, Mariën, & Vandekerckhove, 2014), among others. In social scenarios, where people frequently anticipate others' needs or actions, the understanding of the role of the cerebellum in predictions is particularly relevant (Sokolov, Miall, & Ivry, 2017). Although previous studies (Berthoz, 2002) found increased activity in the cerebellum when predictions (social norms) were violated, we observed the opposite. Hence, our data suggest that in the current context the cerebellum may signal when predictions are matched by social observations. Conversely, when predictions are not met, we observed activation in the lPFC, including the IFG, a region previously associated with semantic cueing (González-García et al., 2016), semantic control (Jefferies, 2013) and emotional regulation during social decisions (Grecucci, Giorgetta, Bonini, & Sanfey, 2013). Therefore, this data also supports the relevance of expectations, even when participants face the outcome of the interaction. At this point, they may need to suppress the previous information to act in accordance with the offer.

Our study has certain limitations, which should be addressed in future investigations. First, the optimal procedure to perform multivariate analyses

and avoid response-related confounds is to counterbalance response options for each participant (Todd, Nystrom, & Cohen, 2013). In the current experiment, however, the association between hand and response was counterbalanced at the group but not the individual level. Thus, our valence-related classification could be confounded by the response patterns linked to acceptance and rejection choices. To rule this out, we performed an additional conjunction analysis between the results of the valence classification and another one performed to classify the response into accept vs. reject choices during the offer period. This conjunction showed that only a small portion of the SMA cluster was common to both contrasts, which indicates that our main valence results could not be explained by the response decoding. In further support of the relevance of the representation of the valance in the bias observed in decisions, our results show that the performance of the classifier for the valence decoding was only related to a specific behavioural bias resulting from the valence of the expectation, not participants' response itself. Therefore, our data highlight that the fidelity of the valence representation in the brain is associated with the extent to which the partners' descriptions modulate participants' decisions.

An additional concern relies on the ecological validity of our study, which is limited by the context of fMRI scanning in a single location. However, we increased the credibility of the social scenario by means of instructions and a cover story, where we recreated an actual delayed interaction between participants of different studies, and where actual earnings were contingent on the choices made during the game. In fact, none of the participants showed signs of susceptibility about the underlying nature of the study when debriefed at the end of the session. Moreover, another step forward would be to assess

participants' personality and prosocial tendencies, since individual predispositions can also influence these dynamics (Díaz-Gutiérrez et al., 2017). Futures studies could use some form of virtual reality during scanning (Mueller et al., 2012) together with more complex verbal descriptions of others to study whether similar brain regions represent this content and the way this is structured. Additionally, another interesting research question would be to try to find if there is a sort of "common valence space" for the two stages of the paradigm. That is, to find out if there is shared information underlying the valence of the adjective (positive/negative) but also the "pleasantness" of the offer (fair-positive, unfair-negative). A future study designed to employ cross-classification decoding approaches (Kaplan, Man, & Greening, 2015) between the expectation and the evidence game periods could advance on this respect.

5.5.1. Conclusions

In this work, we adapted a classical bargaining Ultimatum Game to fMRI, where participants generated positive or negative expectations about partners who afterwards made economic offers. Results highlight the relevance of priors in social decisions, with a set of regions previously related to preparatory processes and cognitive control, semantic cueing, social cognition and valued decisions coding their valence content. Importantly, the fidelity of this information predicted the impact of social priors on posterior choices, stressing the relevance of the patterns of activity in these regions for observable behaviour. Altogether, results suggest that expectations prepare for future decisions, using social information about partners to predict the most likely outcome from the interpersonal exchange.

CHAPTER 6

GENERAL DISCUSSION

The General Discussion is organized into five sections. First, we summarize the main results obtained throughout the Experiments I, II and III. Then, we contextualize our findings within the general framework of control mechanisms and discuss their 'uniqueness' in social scenarios. Afterwards, we discuss some remaining questions and future lines of work, following with the main conclusions that can be drawn from this thesis.

6.1. General summary of results

The main goal of this thesis was to examine the control mechanisms underlying the representation of social information and interference resolution, both in neutral and interpersonal contexts. This way, we have extended the understanding of the representation of social information depending on the task context, as well as the role of control-related networks in social scenarios. To do this, we employed a combination of uni- and multivariate analyses of fMRI data.

In Chapter 3 we directly examined within the same experiment, the common and differential representation of current and intended social task sets in a neutral scenario (Goal 1). For this purpose, we adapted a task-switching paradigm to fMRI where participants had to make sequential social categorization judgements (emotion, gender, race) on human faces. This required that participants maintain these task sets in a currently-active or intended manner. Also, these modifications complement previous work that studied these mechanisms separately and with simpler stimuli (e.g. Woolgar et al., 2011b; Haynes et al., 2007). Univariate analyses supported previous findings associating lateral PFC to the maintenance of task sets and the engagement of frontoparietal regions during task-switching. Importantly, we employed MVPA

to examine fine differences in the representation of social dimensions, depending on when they were relevant (at the moment or later on). Decoding of the current task was possible from orbitofrontal cortex, which together with the fusiform gyrus presented more robust representation of the current vs. intended tasks. This highlights the role of regions associated with social categorization in maintaining the relevant social dimension that is being employed. Conversely, intended task sets were represented in lateral PFC, in line with previous studies with non-social stimuli suggesting that this region represents abstract intentions and goals (Momennejad & Haynes, 2013). Notably, our results showed a region in ventromedial prefrontal cortex that contained information about task sets during either moment, which agrees with a recent proposal in which this region may serve as a map for cognitive states (Schuck et al., 2016).

These results extend previous work that examined the representation of current and intended relevant information to social stimuli. Given the significance of social information in interpersonal scenarios, especially faces, we carried out a second study to examine the influence of social stimuli in interpersonal decisions and the control mechanisms underlying performance in this context (Goal 2). To do this, we employed a Trust Game, where participants had to decide whether to cooperate or not with different partners. We used different cues to guide decisions: social (identity, emotion) and non-social (colour of the frame) dimensions of the presented faces. Importantly, we employed a mixed design (Petersen & Dubis, 2012) to examine the sustained and phasic mechanisms that represent task-relevant information and participate to solve the conflict between the different dimensions. Here we observed general interference effects in behaviour for all dimensions, and specific conflict

markers in frontoparietal regions when emotion and identity were incongruent with each other. This data highlights the relevance of social information in interpersonal decision-making and the role of phasic control in response to conflict between social dimensions that are relevant at different times to predict outcomes.

Despite these results, the complexity of the design hindered a clear interpretation of the findings. Also, it raised the question of how control mechanisms participate to represent social information and their conflict in a different manner in other social scenarios (Goal 3). To this end, we carried out a third study with a different paradigm, the Ultimatum Game. Here we explored proactive and reactive neural markers of the representation of social knowledge in terms of expectations about other people. Also, this paradigm allowed us to examine how expectations interact with actual behaviour, assessing interference markers when participants' expectations were violated. The results from this study showed that social expectations modulate decisions, even when this information is not predictive of people's behaviour. A set of lateral and medial frontal regions increased their activation during the maintenance of the expectations and also contained specific information about the valence (positive or negative) of such predictions. Notably, we observed a relationship between the fidelity of these representations and a behavioural bias, where the better the region distinguished between positive and negative expectations, the more pronounced was participants' tendency to accept an offer after a positive description, compared to their acceptance following a negative expectation. The influence of social expectations was also manifested when participants'

expectations did not match their partners' behaviour, which elicited reactive control by prefrontal mechanisms.

In the following sections we discuss the main findings of the three experiments in terms of how different control mechanisms contribute to the representation of social information and solve their conflict to guide behaviour in neutral and social contexts. Then, we consider the extent to which these mechanisms are specialized for social knowledge or are similar to control in non-social scenarios.

6.2. Active maintenance of social information to represent intentions and expectations.

Our findings bring to light the role of proactive preparation in the maintenance of social information during neutral and interpersonal scenarios. These complement prior work and extend the role of control mechanisms to the representation of social task sets (Experiment I) and the valence of social information during interpersonal decisions (Experiment III).

First, we see that future goals are sustained during ongoing performance in Experiment I. In particular, the representation of intended tasks concurrently to currently-active ones hinders ongoing performance and interfere with current task set information. Importantly, task-relevant information for intended task sets are coded in lateral prefrontal cortex, implying its role not only in storing (Gilbert, 2011) but in the representation of future intentions (Momennejad & Haynes, 2012). Thus, this finding reveals that social task sets can be decoded not only before implementation (Reverberi et al., 2012) during free-delay

periods (Haynes et al., 2007; Momennejad & Haynes, 2013), but also during ongoing performance of other tasks.

In social scenarios, information about others influences decision-making, even when it is not predictive of future outcomes. Although we did not observe any stable maintenance of task sets in Experiment II, we can see in Experiment III that moral descriptions generate expectations about future outcomes that bias decision-making. In addition, such influences take place even when the information provided is not predictive of the actual outcome. These expectations are supported by frontal regions involved in sustained control (Dosenbach et al., 2008) that also code the valence of such predictions. Altogether, these findings extend the role of control mechanisms to the stable maintenance of social information in interpersonal scenarios, and complement previous links between brain and behaviour in non-social contexts (e.g. González-García et al., 2017).

A remarkable finding is that the sustained representation of social information in neutral and interpersonal contexts is supported by close and partially overlapping regions in ventromedial prefrontal cortex (see Figure 1.a). In this line, we see in Experiment I a cluster in medial PFC located ventrally towards the orbitofrontal cortex that maintains a representation of both, current and intended social task sets. Conversely, the larger cluster in Experiment III represents the valence of social expectations during interpersonal decisions. Apart from coding delayed intentions (Gilbert, 2011), different portions of medial prefrontal cortex are associated with specific computations in relation to social cognition (Amodio & Frith, 2006). According to these authors, the orbital

part of prefrontal cortex monitors the value of future outcomes, whereas its anterior part is associated with the linking of value to action and outcomes.

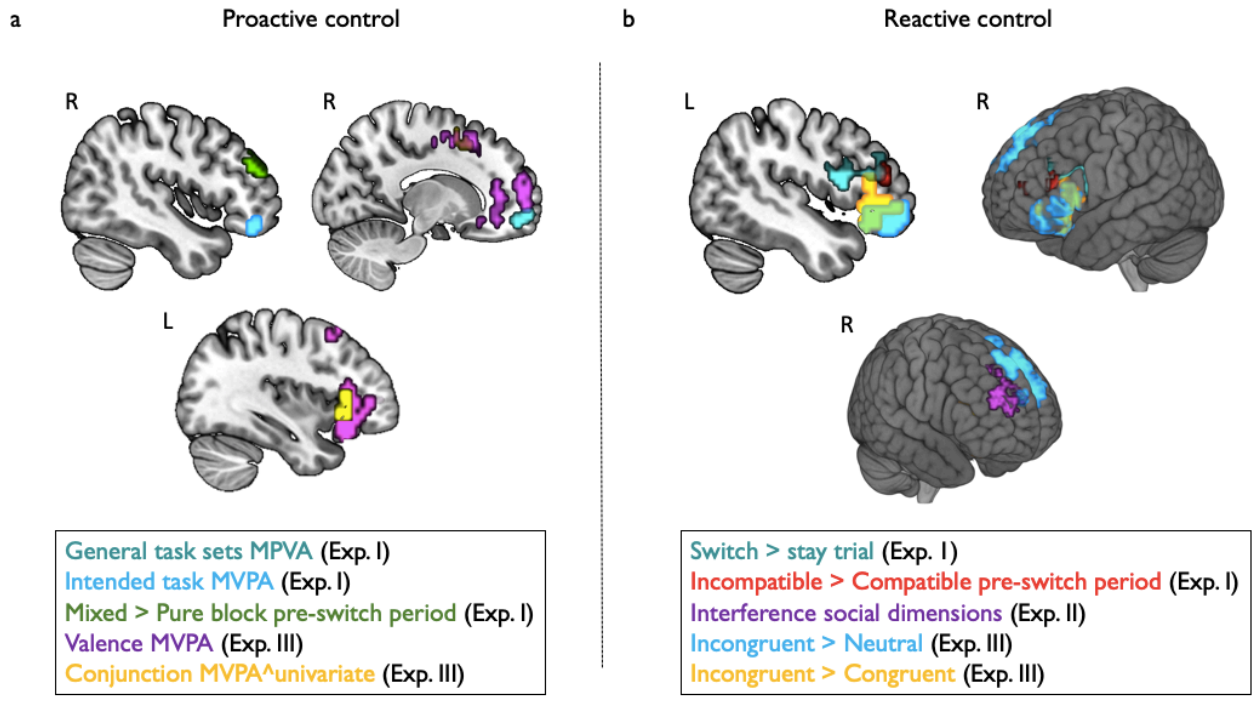

Figure 1. (a) Overlapping across the findings of Experiment I and III for proactive control. (b) Overlapping across the findings of Experiment I, II and III in for reactive control. All the statistical maps are thresholded at $p<.05$ FWE-corrected, except the "Incompatible > Compatible pre-switch period" map, which is shown at a lenient threshold ($p<.001$, $k=22$) for visualization purposes.

Further, Juechems & Summerfield (2019) showed that nearby regions within ventromedial prefrontal cortex represent internal value states during decision-making. In a broader context of decision-making, this region tracks the accumulation of economic resources (Juechems, Balaguer, Ruz, & Summerfield, 2017) and the temporal integration of value (Tsetsos, Wyart, Shorkey, & Summerfield; 2014). Thus, our data from Experiment III complement the role of this region in the representation of social valuation (Bhanji & Delgado, 2014)

of future events. Moreover, these findings are not far from recent proposals (Niv, 2019; Schuck et al., 2016) whereby a close region, medial orbitofrontal cortex, would serve as a map for cognitive states. Altogether, the findings for both Experiments I and III could fit with a role of nearby regions in ventromedial PFC in the representation of sequence of events that take place for a certain behaviour (Juechems et al., 2019), whether they refer to tasks or expected outcomes. Further, our findings complement these proposals with the contribution of ventromedial PFC in the representation of valence during social interactions.

6.3. Reactive control and interference from different sources of social information

Besides the contribution of control mechanisms in the active maintenance of social task-relevant information, we also observe the role of reactive adjustments when the task requires it. We see this first in a neutral scenario during task-switching trials. Thus, we observe in Experiment I that changing from the initial to the intended task recruits lateral prefrontal cortex. The involvement of this region in task-switching is frequently associated with updating of task sets (Sakai, 2008; but see De Baene, Kühn, & Brass, 2012).

In this same context, we see how social task sets based on facial information interfere with each other. In particular, Experiment I showed that the active representation of intended tasks intrudes the performance of the current one, eliciting interference effects when the relevant dimensions for both tasks led to opposite responses. Moreover, Experiment II extends these findings to a social

context and show that frontoparietal regions are recruited when different sources of social information trigger divergent action tendencies.

Likewise, reactive mechanisms also act when the events we are prepared for are not met. Namely, in Experiment III we see that inconsistencies between social expectations about the partners in the game and their actual behaviour recruit prefrontal mechanisms, which could reflect a suppression of previous expectations to decide according to the offer or regulating emotional responses.

Importantly, we see how these different types of conflict engage similar portions of prefrontal cortex across these studies (see Figure 1.b). For that matter, Experiments I and III showed overlapping activation in VLPFC, whereas Experiments II and III recruit DLPFC. Both regions have been associated with reactive control (Palenciano et al., 2017). In particular, the VLPFC underlies rule representation (Crone et al., 2006) and inhibition of competitive task sets (Levy & Wagner, 2011). This is necessary both for task-switching and performance during interference (Experiment I). Also, these computations are important to make decisions when expectations do not match partners' behaviour. Thus, this pattern suggests that for social scenarios the VLPFC would contribute to the retrieval of relevant goals and prior information to guide decisions (Bunge, 2003; Fouragnan et al., 2013). On the other hand, the participation of DLPFC in both social paradigms could indicate the representation of task goals and processing of relevant information during conflict (Egner & Hirsch, 2005; Sanfey et al., 2003). With all, throughout all the experiments we have observed how reactive control is implemented in neutral and social scenarios in response to different type of conflict. Our findings extend

prior work and the recruitment of control mechanisms in this context highlight the impact of social influences in decision-making (Díaz-Gutiérrez et al., 2017; Van Kleef, De Dreu, & Manstead, 2010).

6.4. Control mechanisms for neutral vs. social scenarios

One of the main goals of this thesis was to examine whether control mechanisms have a general role across different contexts or if there are specific control processes that contribute to social scenarios.

In line with control models (Braver, 2012), we show that social information is coded in a sustained fashion prior performance in both neutral and interpersonal scenarios. However, the maintenance and representation of this information are supported by different mechanisms in neutral and social contexts (see Figure 1.a), excepting nearby portions of medial PFC (discussed above). These findings can be due to differences in the material and information participants are maintaining. Thus, in Experiment I participants need to sustain the representation of an intended task set that refers to social categories based on facial stimuli. Conversely, in Experiment III, the information referred to moral descriptions about the partners in the game.

Moreover, interference from social representations triggers reactive adjustments, especially when these are employed in an interpersonal scenario, as we can see in Experiments II and III. Specifically, in neutral and social contexts, the maintenance of social information affects performance, which is evidenced in the interference effects observed in the three experiments. At the neural level, it is remarkable that the findings from the varied studies show

overlapping activations in lateral prefrontal cortex in response to different types of conflict (see Figure 1.b). These patterns suggest that these mechanisms are not tied to the material, since we do not observe overlapping effects between Experiment I and II (both employing faces), but are related to the specificities of the task. This can be observed in the similarities between Experiments I and III, where the preparation for future goals interacts with current performance or actual events, respectively. Also, Experiments II and III are both set in interpersonal contexts where different types of social information interfere with each other: facial dimensions (Experiment II) and expectations vs. partners' behaviour (Experiment III). Further, we wanted to compare if control mechanisms varied depending on the nature of the stimuli (social vs. non-social). In particular, we examined these differences in Experiment II, but we did not find separate activation patterns for the type of relevant information.

Nevertheless, the differences between the paradigms employed in this thesis restrict a direct comparison of findings and limit the scope of our conclusions. Thus, a possible way to overcome this limitation would be developing paradigms as similar as possible where prior instructions could establish the particularities of the context (social or neutral) or the type of social information (faces, descriptions) relevant for task performance. This would provide a more suitable frame for comparison and lead to obtaining sounder conclusions.

Likewise, the lack of larger differences in the contribution of control mechanisms in social vs. neutral contexts as well as for social vs. non-social information could be explained by the paradigm's restrictions in the scanner. The scanning environment prevents participants from a complete immersion in

an interpersonal scenario, which can make participants to engage purely in a computer game where all relevant information is equally processed to guide behaviour. This issue could be mitigated with the use of tools that allow the immersive interaction in more naturalistic environments (Mueller et al., 2012; Reggente et al., 2018), or by employing novel methodologies that allow to transfer experiments out of the lab (as discussed in the next section).

6.5. Remaining questions and future work

Despite the contributions of this thesis, there are still some questions that remain unanswered. First, we examined the preparation for future tasks during ongoing performance, in a sequential judgements' paradigm. However, we lack understanding on how this preparation is organized during the cue period prior to task implementation. This could also be examined in a social setting similar to Experiment II, where we could study the coding of social and non-social task sets during the block's instruction, instead of during block performance. Also, in both cases we could expect to find pre-activations on target-related brain areas (Esterman & Yantis, 2010; González-García et al., 2016), such as specialized face-perception regions (Haxby et al., 2002) and how these preactivations vary depending on the nature of the relevant target (social vs. non-social).

On another note, recent work has developed proposals in relation to the dynamism of person construal (Freeman & Ambady, 2011) and face impressions (Stolier, Hehman, Keller, Walker, & Freeman, 2018), emphasizing the interaction of high and low-level regions in face perception (Stolier & Freeman, 2016). These models suggest that prior conceptual beliefs about others influence how we infer personality traits from face perception. Given the influence of

these processes in social decisions, another important step would be to test these models in interactive scenarios. This would allow to examine how experience during interpersonal decisions shape conceptual trait organization.

The multivariate pattern analyses employed in this thesis provide fine-grained information regarding which regions are informative about the task we are performing, whether they need to focus on the emotion of the person or their race, or if we expect a profitable outcome from an interaction or not. These subtle differences in neural patterns cannot be observed with traditional univariate analysis. For instance, multivariate analyses in Experiment III allowed us to observe regions that are informative about the valence of expectations, which we could not find with the univariate approach. In addition, a recent approach, Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) offers the possibility to study the organization of these representations. Further, RSA allows us to test theoretical models in the brain. For instance, in an adapted version of Experiment II, we could explore if faces were arranged according to the relevant dimension in each block and if this organization was related to performance (i.e. the more the categories are distinguished, the smaller the interference effects).

Moreover, all the experiments in this thesis employed fMRI, which provides information with good spatial resolution. However, this technique lacks a proper temporal resolution to examine in detail the temporal dynamics of control mechanisms, and how the representation of task-relevant information varies through time. Thus, the use of techniques with better temporal resolution such as magnetoencephalography (MEG) or electroencephalography (EEG) will

allow to complement the fMRI data regarding the regions that are informative for specific processes with information about their temporal pattern. In fact, recent studies have combined fMRI and MEG data through RSA (e.g. Hebart, Bankson, Harel, Baker, & Cichy, 2018) to study the spatial and temporal dynamics of task representation.

One of the main goals of this thesis was to examine the influence of social information during social decisions. In Experiment III we included a cover story to increase the credibility of the social interaction, but still, this does not exclude the constraints of the neuroimaging environment (Amodio, 2010). Considering the complexity of social situations, it is necessary to use methodologies that enable us to study these phenomena in more natural contexts, as well as developing more realistic tasks (Gilam & Hendler, 2016). Concerning this, recent studies have employed fNIRS (Pinti et al., 2018) to study different phenomena in natural environments (Mizuno, Hiroyasu, & Hiwa, 2019; Oliver, Tachtsidis, & Hamilton, 2017), which may be a suitable tool to examine the influence of social knowledge in interpersonal contexts.

Last, an ultimate goal of research about social phenomena is to fully understand and characterize interpersonal behaviour, which is present in most dimensions of our lives. In fact, the alteration of social behaviour or social cognition has an impact on multiple areas such as education or health. For instance, impairment of social cognition in one of the manifestations of temporal lobe epilepsy (TLE; Cohn, St-Laurent, Barnett, & McAndrews, 2015; Wang et al., 2015) and frontotemporal dementia (FTD; Gleichgerrcht, Torralva, Roca, Pose, & Manes, 2011), whose patients also show deficits in decision-making (Gleichgerrcht,

Ibanez, Roca, Torralva, & Manes, 2010). This stresses the importance of translating basic research from the lab to other fields, and eventually to society (Cacioppo & Ortigue, 2010). The exploitation of novel methodologies to gather comprehensive knowledge that can be applied to improve the difficulties of those who struggle to navigate social life should be a permanent overarching goal of research in the field.

6.6. Conclusions

Finally, we offer the following conclusions to summarize the present work:

- Social information is actively maintained and represented in both neutral and interpersonal scenarios.
- In neutral contexts, current and intended social task sets are coded in different regions. This way, both high and low-level perceptual regions contain information about currently-active task sets, whereas lateral PFC is in charge of representing abstract intended tasks.
- Ventromedial PFC/OFC maintains a general representation of task-relevant information, regardless of whether it needs to be implemented at the moment or in a delayed manner. This finding agrees with recent work assigning this region a role as a map for cognitive states.
- During social exchanges, information about others influences decision making, even when it is not predictive of the outcome, stressing the relevance of both types of information to guide social decision-making.
- The maintenance of social expectations about partners is supported by frontal regions related to sustained control and preparatory processes. These

regions, together with ventromedial prefrontal cortex, contain specific information about the valence of such expectations.

- The fidelity of the neural coding of social expectations has an impact on subsequent decisions, where the strength of the neural representation of valence predicted the extent to which participants are influenced by social information to make decisions.
- Interference from different sources of social information declines performance and triggers adjustments in frontoparietal areas, mainly during social interactions.
- Importantly, different types of conflict across the studies engage overlapping portions of lateral PFC, suggesting their role underlying common reactive computations for neutral and interpersonal contexts.

RESUMEN EN CASTELLANO

Dentro de las habilidades del ser humano destaca nuestra flexibilidad y capacidad de adaptación al medio, donde estamos expuestos a numerosos contextos y fuentes de información de distinta naturaleza, ante los cuales hemos de ser capaces de proporcionar una conducta apropiada. Esto es especialmente necesario en situaciones de cierta dificultad, como cuando procedimientos rutinarios conducen a respuestas divergentes (Norman & Shallice, 1980). Estas habilidades, también conocidas como control cognitivo, nos permiten navegar de forma exitosa por un entorno cambiante y complejo. Los distintos estudios sobre los mecanismos de control coinciden en la importancia de regiones fronto-parietales que cumplen un rol en el ajuste de nuestra conducta en contextos de alta demanda cognitiva, también llamadas conjuntamente red de múltiple demanda (MD, por sus siglas en inglés; Duncan, 2010). Estas regiones contribuyen cuando es necesario preparar un set de tarea (control proactivo), así como para ejercer ajustes ante la presentación de estímulos (control reactivo, Braver, 2012). Por otro lado, Dosenbach et al., (2008) propusieron una distinción de los mecanismos de control en redes neurales con perfiles complementarios, puestos en marcha durante tareas complejas que requieren esfuerzo. Así, una red fronto-parietal se pondría en marcha de forma fásica, como respuesta a eventos, mientras que una red cíngulo-opercular mantendría una actividad sostenida con el fin de representar el contexto de tarea de forma estable. Con todo lo anterior, cabe destacar que, si bien el trabajo acerca de los mecanismos de control es extenso, estos se centran en el empleo de estímulos y paradigmas simples (aunque ver González-García et al., 2017; Palenciano et al., 2019a,b), o ajenos a un contexto social.

Dado que las personas nos vemos expuestas constantemente a interacciones sociales, donde continuamente múltiples fuentes de información personal generan expectativas acerca del comportamiento o pensamiento de los demás, se entiende que este tipo de procesos van ligados también a cierto esfuerzo. Como muestra, diversos estudios han explorado los aspectos estratégicos del comportamiento durante estas interacciones sociales (e.g. Ruz et al. 2013), ya que, aunque nuestras decisiones interpersonales estén acompañadas de una sensación de objetividad, estas se encuentran sesgadas por información evaluativa acerca de los otros (Díaz-Gutiérrez et al. 2017), aunque la misma no tenga relación con su comportamiento (Alguacil et al., 2015; Tortosa et al., 2013).

Con todo esto, la presente tesis ha tenido como objetivo el estudio de las redes mencionadas en el mantenimiento y representación de información social y su interferencia en contextos neutros e interpersonales. Para responder a este objetivo, hemos realizado tres experimentos aplicando técnicas uni y multivariadas a datos de resonancia magnética funcional (RMf).

En primer lugar, examinamos en un contexto neutro cómo la representación de estímulos sociales varía en función de cuando sea necesario emplear dicha información. Con este objetivo, empleamos un paradigma de cambio de tarea adaptado a RMf, donde los participantes tenían que categorizar dimensiones sociales de estímulos faciales (raza, género o emoción) y necesitaban mantener los sets de tarea de forma actual o demorada. Asimismo, empleamos análisis de patrones multivariados (MVPA, por sus siglas en inglés) para estudiar diferencias finas en la representación neural de estas dimensiones sociales,

dependiendo de cuando fueran relevantes (actualmente o de forma demorada). Los resultados muestran que los sets de tarea demorados se representan de manera diferente a la tarea que se está realizando en el momento. De esta forma, la corteza orbitofrontal lateral es informativa de la tarea actual en solitario, mientras que la información demorada se representa en la corteza prefrontal lateral. Este resultado apoya estudios previos que indican que esta región codifica sets de tarea demorados cuando estos tienen cierto grado de abstracción (Momennejad & Haynes, 2013). Asimismo, entre los hallazgos de este estudio destaca que la corteza prefrontal ventromedial contiene información de las tareas relevantes independientemente de cuando se emplee esta información (ahora o de forma demorada). Este resultado se encuentra en línea con una propuesta reciente por la cual esta región serviría como un mapa de estados cognitivos (Schuck et al., 2016).

Dada la importancia que tiene la información social, y en concreto, los rostros humanos para predecir el comportamiento de las personas, decidimos estudiar los mecanismos de control durante un contexto interpersonal. En este caso adaptamos un juego de confianza (Trust Game; Camerer, 2003) donde los participantes debían decidir si cooperar con una serie de compañeros, y donde el resultado final del intercambio dependía de la reciprocidad entre el participante y los compañeros de juego. Las distintas fuentes de información utilizadas permitían estudiar los mecanismos de control en contextos sociales, así como comparar cuando se utilizan dimensiones sociales (identidad personal o expresión emocional) o no sociales (color de un marco) para realizar predicciones acerca de la conducta de la otra persona. Para ello, empleamos un diseño híbrido (Petersen & Dubis, 2012; Visscher et al., 2003) para estudiar los

mecanismos sostenidos y fásicos que representan esta información. En este caso, observamos marcadores generales de conflicto entre las distintas fuentes de información. Aunque no encontramos evidencia del mantenimiento sostenido de las dimensiones relevantes en cada bloque, la interferencia entre dimensiones sociales reclutó regiones fronto-parietales de control de manera fásica, lo que realza la importancia de la información social en la toma de decisiones interpersonal.

Por último, quisimos trasladar el estudio de estos mecanismos a otro contexto interpersonal donde la información social provenga de otra fuente. De este modo, adaptamos un paradigma diferente de toma de decisiones sociales, en este caso el Ultimatum Game (Güth, 1982). En este caso quisimos explorar los mecanismos neurales subyacentes a la codificación de información social y generación de expectativas de otras personas. Dicho paradigma nos permitió examinar la interacción de dichas expectativas y la conducta de los compañeros de juego. Los resultados de dicho estudio muestran cómo las expectativas modulan las decisiones de los participantes, aun sin ser predictivas del comportamiento del compañero. Asimismo, una serie de regiones frontales se encargan de representar, de forma preparatoria, información sobre las características morales de los compañeros de juego. Además, el uso de MVPA nos permitió decodificar la valencia de dichas expectativas en esas mismas regiones, así como en la corteza prefrontal medial. Cabe destacar que cuanto mejor era capaz cada región de diferenciar si la información sobre el compañero era positiva o negativa, mayor era el sesgo del participante de dejarse llevar por esta información para tomar su decisión. Por otro lado, cuando estas

expectativas no son seguidas del comportamiento esperado de los compañeros, entran en acción regiones prefrontales de control cognitivo.

En conjunto, nuestros resultados destacan la versatilidad de los mecanismos de control. Estos subyacen a la preparación para tareas futuras, así como el mantenimiento de expectativas sobre otras personas. Asimismo, la representación de información social afecta al desempeño y a la toma de decisiones interpersonales. De esta forma, se requieren ajustes por parte de mecanismos reactivos cuando distintas fuentes de información conducen a respuestas opuestas o cuando la información previa nos conduce a unas predicciones que no son posteriormente cumplidas.

ABSTRACT

The ability to behave flexibility and in adaptation to the environment are among the repertoire of human skills. We are constantly exposed to numerous contexts and sources of information, to which we need to provide a suitable response. This is especially necessary when we face situations of certain complexity, such as when routine procedures lead to divergent responses (Norman & Shallice, 1980). This set of skills, better known as cognitive control, allows us to navigate changing and challenging situations. The studies on control mechanisms agree on the importance of frontoparietal regions that implement adjustments in high demand scenarios, which have been jointly named as the Multiple Demand Network (MD; Duncan, 2010). These regions contribute when it is necessary to prepare a task set (proactive control) as well as to provide responses at stimuli presentation (reactive control, Braver, 2012). Further, Dosenbach et al. (2008) proposed a distinction of control mechanisms in different networks with complementary profiles, initiated during effortful tasks. Thus, a frontoparietal network would act transiently in response to events, whereas a cinguloopercular one would sustain an active and stable representation of task context. Although the work on control mechanisms is extensive, most studies focus on stimuli or paradigms rather simple (but see Egner et al., 2008; González-García et al., 2017; Palenciano et al., 2019a,b) or lacking social context.

As social beings, humans are exposed constantly to interactions with other people, where multiple sources of personal information generate expectations about them. This suggests that these phenomena are tied to some effort. In fact, several studies have explored strategic behaviour during social interactions (e.g. Ruz et al., 2013), since even if our interpersonal decisions are followed by a sense of rationality, these are biased by evaluative information about others

(Díaz-Gutiérrez et al., 2017), even if it is unrelated to their behaviour (Alguacil et al., 2015; Tortosa et al., 2013).

With all of the above, this thesis has aimed to study the role of control mechanisms in the maintenance and representation of social information and its interference in neutral and interpersonal contexts. To fulfil this goal, we have carried out three experiments combining with uni- and multivariate analyses on fMRI data.

First, we aimed to examine, in a neutral context, how the representation of social stimuli varied depending on when it is necessary to implement such information. To do this, we employed a task-switching paradigm adapted to fMRI, where participants had to make sequential categorization judgements on human faces (race, gender, emotion) and needed to maintain these task sets in a currently-active or delayed manner. In addition, we used MPVA to study fine-grained differences in the neural representation of social dimensions, depending on when they were relevant. Results show that intended task sets are represented differently from the currently-active task. Thus, lateral orbitofrontal cortex codes the current task only, whereas lateral prefrontal cortex represents intended task sets. This finding supports previous studies indicating that this region contains specific information about intended abstract goals (Momennejad & Haynes, 2013). Moreover, a remarkable finding is that a region located in ventromedial prefrontal cortex/orbitofrontal cortex contains information about relevant task sets regardless of when they need to be implemented (now or later on), in line with a recent proposal whereby this region would serve as a map for cognitive states (Schuck et al., 2016).

Given the relevance of social information, and in particular, human faces to predict people's behaviour, we decided to examine control mechanisms in an interpersonal context. Here, we adapted a Trust Game (Camerer, 2003) where participants needed to decide to cooperate or not with a series of partners and the final outcome from the interaction depended on reciprocation between the participant and the partners in the game. Further, different dimensions could predict partners behaviour, so we could compare the role of control when this information was social (identity or emotional expression) or non-social (colour of a frame). On top of that, we employed a mixed design (Petersen & Dubis, 2012; Visscher et al., 2003) to study sustained and transient mechanisms in charge of representing social and non-social relevant dimensions and their conflict. Results show general behavioural interference effects between the different dimensions. Also, conflict between the social dimensions elicited transient activation in frontoparietal regions, which highlights the relevance of social information in interpersonal decisions.

Last, we examined control mechanisms in another interpersonal context in which social information came from descriptions about the partners' moral traits. This way, we adapted a different paradigm, an Ultimatum Game (Güth, 1982). In this case, we wanted to explore the neural mechanisms underlying the representation of social information and the generation of expectations about others. Likewise, this paradigm allowed us to examine the interaction between such predictions and partners' actual behaviour. The findings from this study showed how social expectations modulate decisions, even if they are not predictive of people's behaviour. Moreover, a set of frontal regions represented

the valence of participants' expectations. Importantly, the better these areas represented this information, the more biased were participants' decisions by it. Further, the influence of social information was manifested when participants' expectations did not match their partners' behaviour, which triggered frontal activation in control-related areas.

Overall, our findings highlight the versatility of control mechanisms. These underlie the preparation for future tasks, as well as the maintenance of social expectations. Further, the representation of social information affects performance and interpersonal decisions making. In this way, reactive mechanisms are required to implement appropriate adjustments when different sources of information lead to divergent responses or when social information induces predictions about others that are not met.

REFERENCES

Abdulrahman, H., & Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. *NeuroImage*, *125*, 756–766. https://doi.org/10.1016/j.neuroimage.2015.11.009

Adolphs, R., & Anderson, D. (2013). Social and emotional neuroscience. *Current Opinion in Neurobiology*, *23*(3), 291–293. https://doi.org/10.1016/j.conb.2013.04.011

Alguacil, S., Díaz-Gutiérrez, P., Kotz, S. A., Mestres-Misse, A., Tudela, P. & Ruz, M. (2016). *Emotional conflict between personal identity and emotional expression: an fMRI study*. Unpsublished manuscript.

Alguacil, S., Madrid, E., Espín, A. M., & Ruz, M. (2017). Facial identity and emotional expression as predictors during economic decisions. *Cognitive, Affective and Behavioral Neuroscience*, *17*(2), 315–329. https://doi.org/10.3758/s13415-016-0481-9

Alguacil, S., Tudela, P., & Ruz, M. (2015). Ignoring facial emotion expressions does not eliminate their influence on cooperation decisions. *Psicologica*, *36*(2), 309–335.

Amodio, D. M. (2010). Can Neuroscience advance Social Psychology Theory? Social Neuroscience for the Behavioral Social Pyschologist. *Social Cognition*, *28*(6), 695–716.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Apps, M. A. J., Lockwood, P. L., & Balsters, J. H. (2013). The role of the midcingulate cortex in monitoring others' decisions. *Frontiers in Neuroscience*, *7*, 251. https://doi.org/10.3389/fnins.2013.00251

Arco, J. E., González-García, C., Díaz-Gutiérrez, P., Ramírez, J., & Ruz, M. (2018). Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis. *Journal of Neuroscience Methods*, *308*, 248–260. https://doi.org/10.1016/j.jneumeth.2018.06.017

Baddeley, A. D. (1992). Working Memory. *Science*, *255*(5044), 556–559.

Bar, A. M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primative. *Advances in Experimental Social Psychology*, *41*: 167–218. https://doi.org/10.1016/S0065-2601(08)00404-8.

Berthoz, S. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain*, *125*(8), 1696–1708. https://doi.org/10.1093/brain/awf190

Bhandari, A., Gagne, C., & Badre, D. (2018). Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? *Journal of Cognitive Neuroscience*, *30*(10), 1473–1498. https://doi.org/10.1162/jocn

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 61–73. https://doi.org/10.1002/wcs.1266

Bode, S., & Haynes, J. D. (2009). Decoding sequential stages of task preparation in the human brain. *NeuroImage*, *45*(2), 606–613. https://doi.org/10.1016/j.neuroimage.2008.11.031

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037//0033-295X.I08.3.624

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports*, *4*, 6025. https://doi.org/10.1038/srep06025

Brass, M., & von Cramon, D. Y. (2002). The Role of the Frontal Cortex in Task Preparation. *Cerebral Cortex*, *12*(9), 908–914. https://doi.org/10.1093/cercor/12.9.908

Brass, M., & von Cramon, D. Y. (2004). Decomposing Components of Task Preparation with Functional Magnetic Resonance Imaging. *Journal of Cognitive Neuroscience*, *16*(4), 609–620. https://doi.org/10.1162/089892904323057335

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Reynolds, J. R., & Donaldson, D. I. (2003). Neural Mechanisms of Transient and Sustained Cognitive Control during Task Switching. *Neuron*, *39*(4), 713–726. https://doi.org/10.1016/S0896-6273(03)00466-5

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The Brain's Default Network. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in*

Cognitive Sciences, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective and Behavioral Neuroscience*, *4*(4), 564–579. https://doi.org/10.3758/CABN.4.4.564

Burgess, P. W., Scott, S. K., & Frith, C. D. (2003). The role of the rostral frontal cortex (area 10) in prospective memory: A lateral versus medial dissociation. *Neuropsychologia*, *41*(8), 906–918. https://doi.org/10.1016/S0028-3932(02)00327-5

Cacioppo, J. T., & Ortigue, S. (2010). Social Neuroscience : How a Multidisciplinary Field Is Uncovering the Biology of Human Interactions. *Cerebrum*, (December), 1–12.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interactions*. Princeton: Princeton University Press.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*(3), 277–284. https://doi.org/10.1093/scan/nsr094

Cohn, M., St-Laurent, M., Barnett, A., & McAndrews, M. P. (2015). Social inference deficits in temporal lobe epilepsy and lobectomy: Risk factors and neural substrates. *Social Cognitive and Affective Neuroscience*, *10*(5), 636–644. https://doi.org/10.1093/scan/nsu101

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*, *7*(7), 764–770. https://doi.org/10.1093/scan/nsr053

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel Patterns in Fusiform Face Area Differentiate Faces by Sex and Race. *PLoS ONE*, *8*(7), e69684. https://doi.org/10.1371/journal.pone.0069684

Corradi-Dell'Acqua, C., Turri, F., Kaufmann, L., Clément, F., & Schwartz, S. (2015). How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation. *Cortex*, *70*, 21–34. https://doi.org/10.1016/j.cortex.2015.02.011

Correa, A., Ruiz-Herrera, N., Ruz, M., Tonetti, L., Martoni, M., Fabbri, M., & Natale, V. (2017). Economic decision-making in morning/evening-type people as a function of time of day. *Chronobiology International*, *34*(2), 139–147. https://doi.org/10.1080/07420528.2016.1246455

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1(1)*, 42-45. https://doi.org/10.20982/tqmp.01.1.p042

Crittenden, B. M., Mitchell, D. J., & Duncan, J. (2015). Recruitment of the default mode network during a demanding act of executive control. *ELife*, *4:e06481*. https://doi.org/10.7554/eLife.06481

Crittenden, B. M., Mitchell, D. J., & Duncan, J. (2016). Task encoding across the multiple demand cortex is consistent with a frontoparietal and cingulo-opercular dual networks distinction. *Journal of Neuroscience*, *36*(23), 6147–6155. https://doi.org/10.1523/JNEUROSCI.4590-15.2016

Crone, E. A., Wendelken, C., Donohue, S. E., & Bunge, S. A. (2006). Neural evidence for dissociable components of task-switching. *Cerebral Cortex*, *16*(4), 475–486. https://doi.org/10.1093/cercor/bhi127

De Baene, W., & Brass, M. (2014). Dissociating strategy-dependent and

independent components in task preparation. *Neuropsychologia*, *62*, 331–340. https://doi.org/10.1016/j.neuropsychologia.2014.04.015

De Baene, W., Kühn, S., & Brass, M. (2012). Challenging a decade of brain research on task switching: Brain activation in the task-switching paradigm reflects adaptation rather than reconfiguration of task sets. *Human Brain Mapping*, *33*(3), 639–651. https://doi.org/10.1002/hbm.21234

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.neuro.18.1.193

Díaz-Gutiérrez, P., Alguacil, S., & Ruz, M. (2017). Bias and control in social decision-making. In A. Ibáñez, L. Sedeño, & A. Gacría (Eds.), *Neuroscience and Social Science: The Missing Link* (pp. 47–68). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-68421-5

Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in Cognitive Sciences*, *12*(3), 99–105. https://doi.org/10.1016/j.tics.2008.01.001

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., Fox, M. D., Snyder, A.Z., Vincent, J.L., Raichle, ME, Schlaggar, B. L. & Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences*, *104*(26), 11073–11078. https://doi.org/10.1073/pnas.0704320104

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A.L., Schlaggar, B.L., Petersen, S. E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, *50*(5), 799–812. https://doi.org/10.1016/j.neuron.2006.04.031

Dubis, J. W., Siegel, J. S., Neta, M., Visscher, K. M., & Petersen, S. E. (2016). Tasks Driven by Perceptual Information Do Not Recruit Sustained BOLD Activity in Cingulo-Opercular Regions. *Cerebral Cortex*, *26*(1), 192–201. https://doi.org/10.1093/cercor/bhu187

Dumontheil, I., Thompson, R., & Duncan, J. (2011). Assembly and Use of New Task Rules in Fronto-parietal Cortex. *Journal of Cognitive Neuroscience*, *23*(1), 168–182. https://doi.org/10.1162/jocn.2010.21439

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*(10), 374–380. https://doi.org/10.1016/j.tics.2008.07.001

Egner, T., Etkin, A., Gale, S., & Hirsch, J. (2008). Dissociable neural systems resolve conflict from emotional versus nonemotional distracters. *Cerebral Cortex*, *18*, 1475–1484. https://doi.org/10.1093/cercor/bhm179

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784–1790. https://doi.org/10.1038/nn1594

Elton, A., & Gao, W. (2015). Task-positive Functional Connectivity of the Default Mode Network Transcends Task Domain. *Journal of Cognitive Neuroscience*, *27*(12), 2369–2381. https://doi.org/10.1162/jocn

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Esterman, M., & Yantis, S. (2010). Perceptual expectation evokes category-selective cortical activity. *Cerebral Cortex*, *20*(5), 1245–1253. https://doi.org/10.1093/cercor/bhp188

Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving Emotional Conflict: A Role for the Rostral Anterior Cingulate Cortex in Modulating Activity in the Amygdala. *Neuron*, *51*(6), 871–882. https://doi.org/10.1016/j.neuron.2006.07.029

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience*, *35*(21), 8170–8180. https://doi.org/10.1523/JNEUROSCI.4775-14.2015

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427. https://doi.org/10.1016/j.tics.2007.09.002

Filkowski, M. M., Anderson, I. W., & Haas, B. W. (2016). Trying to trust: Brain activity during interpersonal social attitude change. *Cognitive, Affective and Behavioral Neuroscience*, *16*(2), 325–338. https://doi.org/10.3758/s13415-015-0393-0

Fleming, S. M., Thomas, C. L., & Dolan, R. J. (2010). Overcoming status quo bias in the human brain. *Proceedings of the National Academy of Sciences*, *107*(13), 6005–6009. https://doi.org/10.1073/pnas.0910380107

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *The Journal of Neuroscience*, *33*(8), 3602–3611.

https://doi.org/10.1523/JNEUROSCI.3086-12.2013

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. Van, & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678. https://doi.org/10.1073/pnas.0504136102

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

Freeman, J. B., Ma, Y., Barth, M., Young, S. G., Han, S., & Ambady, N. (2015). The neural basis of contextual influences on face categorization. *Cerebral Cortex*, *25*(2), 415–422. https://doi.org/10.1093/cercor/bht238

Freeman, J. B., Rule, N. O., Adams, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*(6), 1314–1322. https://doi.org/10.1093/cercor/bhp195

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Frith, C. D. (2007). The social brain? *Phil. Trans. R. Soc. B*, *362*(10), 671–678. https://doi.org/10.1098/rstb.2006.2003

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*(1), 36–46. https://doi.org/10.1016/j.brainres.2005.12.126

Frith, C. D., & Frith, U. (2008). Implicit and Explicit Processes in Social Cognition. *Neuron*, *60*(3), 503–510.

https://doi.org/10.1016/j.neuron.2008.10.032

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *American Psychological Society*, *10*(5), 151–155. https://doi.org/https://doi.org/10.1111/1467-8721.00137

Gabay, A. S., Radua, J., Kempton, M. J., & Mehta, M. A. (2014). The Ultimatum Game and the brain: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *47*, 549–558. https://doi.org/10.1016/j.neubiorev.2014.10.014

Gaertig, C., Moser, A., Alguacil, S., & Ruz, M. (2012). Social information and economic decision-making in the ultimatum game. *Frontiers in Neuroscience*, *6:103*. https://doi.org/10.3389/fnins.2012.00103

Gilam, G., & Hendler, T. (2016). With love, from me to you: Embedding social interactions in affective neuroscience. *Neuroscience and Biobehavioral Reviews*, *68*, 590–601. https://doi.org/10.1016/j.neubiorev.2016.06.027

Gilbert, S. J. (2011). Decoding the Content of Delayed Intentions. *Journal of Neuroscience*, *31*(8), 2888–2894. https://doi.org/10.1523/JNEUROSCI.5336-10.2011

Gilbert, S. J., & Fung, H. (2018). Decoding intentions of self and others from fMRI activity patterns. *NeuroImage*, *172*, 278–290. https://doi.org/10.1016/j.neuroimage.2017.12.090

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*(14), 3600–3611. https://doi.org/10.1016/j.neuropsychologia.2012.09.002

Gleichgerrcht, E, Ibanez, A., Roca, M., Torralva, T., & Manes, F. (2010). Decision-making cognition in neurodegenerative diseases. *Nat Rev Neurol*,

6(11), 611–623. https://doi.org/10.1038/nrneurol.2010.148

Gleichgerrcht, E., Torralva, T., Roca, M., Pose, M., & Manes, F. (2011). The role of social cognition in moral judgment in frontotemporal dementia. *Social Neuroscience*, *6*(2), 113–122. https://doi.org/10.1080/17470919.2010.506751

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., & Ruz, M. (2017). Encoding, preparation and implementation of novel complex verbal instructions. *NeuroImage*, *148*(January), 264–273. https://doi.org/10.1016/j.neuroimage.2017.01.037

González-García, C., Mas-Herrero, E., de Diego-Balaguer, R., & Ruz, M. (2016). Task-specific preparatory neural activations in low-interference contexts. *Brain Structure and Function*, *221*(8), 3997–4006. https://doi.org/10.1007/s00429-015-1141-5

Gratton, C., Neta, M., Sun, H., Ploran, E. J., Schlaggar, B. L., Wheeler, M. E., Petersen, S. E. & Nelson, S. M. (2017). Distinct Stages of Moment-to-Moment Processing in the Cinguloopercular and Frontoparietal Networks. *Cerebral Cortex*, *27*(3), 2403–2417. https://doi.org/10.1093/cercor/bhw092

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: the role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience*, *7:523*. https://doi.org/10.3389/fnhum.2013.00523

Greene, J., & J, H. (2002). How (and where) does moral judgement work? *Trends in Cognitive Sciences*, *6*(12), 517–523. https://doi.org/10.1016/s1364-6613(02)02011-9

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of

ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *NeuroImage*, *50*(3), 1313–1319. https://doi.org/10.1016/j.neuroimage.2009.12.109

Hampson, M., Driesen, N. R., Skudlarski, P., Gore, J. C., & Constable, R. T. (2006). Brain Connectivity Related to Working Memory Performance. *Journal of Neuroscience*, *26*(51), 13338–13343. https://doi.org/10.1523/JNEUROSCI.3408-06.2006

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, *24*(8), 1979–1987. https://doi.org/10.1093/cercor/bht042

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The Distributed Human Neural System for Face Perception. *Trends in Cognitive Sciences, 4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems

for face recognition and social communication. *Biological Psychiatry*, *51*(1), 59–67. https://doi.org/10.1016/S0006-3223(01)01330-0

Haynes, J. D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, *17*(4), 323–328. https://doi.org/10.1016/j.cub.2006.11.072

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *ELife*, *7*, 1–21. https://doi.org/10.7554/eLife.32816

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Henson, R. N., Goshen-Gottstein, Y., Ganel, T., Otten, L. J., Quayle, A., & Rugg, M. D. (2003). Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cerebral Cortex*, *13*(7), 793–805. https://doi.org/10.1093/cercor/13.7.793

Humphrey, N. (1976). The social function of intellect. In P. Bateson. and R.A. Hinde (EdS.), *Growing Points in Ethology* (pp. 303–317). Cambridge: Cambridge University Press.

Jefferies, E. (2013). The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. *Cortex*, *49*(3), 611–625. https://doi.org/10.1016/j.cortex.2012.10.008

Juechems, K., Balaguer, J., Herce Castañón, S., Ruz, M., O'Reilly, J. X., &

Summerfield, C. (2019). A Network for Computing Value Equilibrium in the Human Medial Prefrontal Cortex. *Neuron*, *101*(5), 977-987.e3. https://doi.org/10.1016/j.neuron.2018.12.029

Juechems, K., Balaguer, J., Ruz, M., & Summerfield, C. (2017). Ventromedial Prefrontal Cortex Encodes a Latent Estimate of Cumulative Reward. *Neuron*, *93*(3), 705-714.e4. https://doi.org/10.1016/j.neuron.2016.12.038

Juechems, K., & Summerfield, C. (2019). Where Does Value Come From? *Trends in Cognitive Sciences*, *23*(10), 836–850. https://doi.org/10.1016/j.tics.2019.07.012

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and Assumptions of Economics. *Journal of Business*, *59*(4), S285-300.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience,* 17(11), 4302–4311. https://doi.org/10.1109/CDC.2005.1583375

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience*, *9*, 151. https://doi.org/10.3389/fnhum.2015.00151

Kaul, C., Ratner, K. G., & Van Bavel, J. J. (2014). Dynamic representations of race: Processing goals shape race decoding in the Fusiform gyri. *Social Cognitive and Affective Neuroscience*, *9*(3), 326–332.

https://doi.org/10.1093/scan/nss138

Kaul, C., Rees, G., & Ishai, A. (2011). The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain. *Frontiers in Human Neuroscience*, 4:238. https://doi.org/10.3389/fnhum.2010.00238

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, *303*(5660), 1023–1026. https://doi.org/10.1126/science.1089910

Kliegel, M., McDaniel, M., & Einstein, G. (2008). *Prospective memory: cognitive, neuroscience, developmental, and applied perspectives*. Mahwah, NJ: Erlbaum.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829–832. https://doi.org/10.1126/science.1129156

Koski, J. E., Collins, J. A., Olson, I. R., & Hospital, G. (2017). The neural representation of social status in the extended face- processing network. *European Journal of Neuroscience*, *46*(12), 2795–2806. https://doi.org/10.1111/ejn.13770.The

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity

analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2: 4*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Landsiedel, J., & Gilbert, S. J. (2015). Creating external reminders for delayed intentions: Dissociable influence on "task-positive" and "task-negative" brain networks. *NeuroImage*, *104*, 231–240. https://doi.org/10.1016/j.neuroimage.2014.10.021

Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience*, *7*, 259. https://doi.org/10.3389/fnins.2013.00259

Lesage, E., Hansen, P. C., & Miall, R. C. (2017). Right Lateral Cerebellum Represents Linguistic Predictability. *The Journal of Neuroscience*, *37*(26), 6231–6241. https://doi.org/10.1523/JNEUROSCI.3203-16.2017

Levens, S. M., & Phelps, E. A. (2010). Insula and orbital frontal cortex activity underlying emotion interference resolution in working memory. *Journal of Cognitive Neuroscience*, *22*(12), 2790–2803. https://doi.org/10.1162/jocn.2010.21428

Levy, B. J., & Wagner, A. D. (2011). Cognitive control and right ventrolateral prefrontal cortex: Reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences*, *1224*(1), 40–62. https://doi.org/10.1111/j.1749-6632.2011.05958.x

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., & Barrett, L. F. (2015). The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. *Cerebral Cortex*, *26*(5),

1910–1922. https://doi.org/10.1093/cercor/bhv001

Loose, L. S., Wisniewski, D., Rusconi, M., Goschke, T., & Haynes, J.-D. (2017). Switch-Independent Task Representations in Frontal and Parietal Cortex. *The Journal of Neuroscience*, *37*(33), 8033–8042. https://doi.org/10.1523/jneurosci.3656-16.2017

Lopez-Persem, A., Domenech, P., & Pessiglione, M. (2016). How prior preferences determine decision-making frames and biases in the human brain. *ELife*, *5: e20317*. https://doi.org/10.7554/eLife.20317

Los, S. A. (1996). On the origin of mixing costs: Exploring information processing in pure and mixed blocks of trials. *Acta Psychologica*, *94*(2), 145–188. https://doi.org/10.1016/0001-6918(95)00050-X

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. Van, Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*(8), 937–950. https://doi.org/10.1093/scan/nsr064

Marí-Beffa, P., Cooper, S., & Houghton, G. (2012). Unmixing the mixing cost: Contributions from dimensional relevance and stimulus-response suppression. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 478–488. https://doi.org/10.1037/a0025979

Marini, F., Demeter, E., Roberts, K. C., Chelazzi, L., & Woldorff, M. G. (2016). Orchestrating proactive and reactive mechanisms for filtering distracting information: Brain-behavior relationships revealed by a mixed-design fMRI study. *Journal of Neuroscience*, *36*(3), 988–1000. https://doi.org/10.1523/JNEUROSCI.2966-15.2016

Mars, R. B., Neubert, F. X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. (2012). On the relationship between the “default mode network” and the

"social brain." *Frontiers in Human Neuroscience*, *6*, 1–9. https://doi.org/10.3389/fnhum.2012.00189

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Meyer, M. L. (2019). Social by Default: Characterizing the Social Functions of the Resting Brain. *Current Directions in Psychological Science, 28*(4)*, 380-386*. https://doi.org/10.1177/0963721419857759

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2018). Evidence That Default Network Connectivity During Rest Consolidates Social Information. *Cerebral Cortex*, *29*(5), 1910-1920. https://doi.org/10.1093/cercor/bhy071

Meyer, M. L., & Lieberman, M. D. (2016). Social Working Memory Training Improves Perspective-Taking Accuracy. *Social Psychological and Personality Science*, *7*(4), 381–389. https://doi.org/10.1177/1948550615624143

Miller, E. K., & Cohen, J. D. (2001). An integrate theory of PFC function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron*, *50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, *320*, 1191–1195. https://doi.org/10.1126/science.1152876

Mizuno, M., Hiroyasu, T., & Hiwa, S. (2019). A functional NIRS study of brain functional networks induced by social time coordination. *Brain Sciences*, *9*(2), 1–11. https://doi.org/10.3390/brainsci9020043

Momennejad, I., & Haynes, J. D. (2012). Human anterior prefrontal cortex encodes the "what" and "when" of future intentions. *NeuroImage*, *61*(1), 139–148. https://doi.org/10.1016/j.neuroimage.2012.02.079

Momennejad, I., & Haynes, J. D. (2013). Encoding of Prospective Tasks in the Human Prefrontal Cortex under Varying Task Loads. *Journal of Neuroscience*, *33*(44), 17342–17349. https://doi.org/10.1523/JNEUROSCI.0492-13.2013

Monsell, S. (2003). Task Switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/doi:10.1016/S1364-6613(03)00028-7

Morgan, H. M., Jackson, M. C., Van Koningsbruggen, M. G., Shapiro, K. L., & Linden, D. E. J. (2013). Frontal and parietal theta burst TMS impairs working memory for visual-spatial conjunctions. *Brain Stimulation*, *6*(2), 122–129. https://doi.org/10.1016/j.brs.2012.03.001

Moser, A., Gaertig, C., & Ruz, M. (2014). Social information and personal interests modulate neural activity during economic decision-making. *Frontiers in Human Neuroscience*, *8: 31*. https://doi.org/10.3389/fnhum.2014.00031

Moyes, J., Sari-Sarraf, N., & Gilbert, S. J. (2019). Characterising monitoring processes in event-based prospective memory: Evidence from pupillometry. *Cognition*, *184*, 83–95. https://doi.org/10.1016/j.cognition.2018.12.007

Mueller, C., Luehrs, M., Baecke, S., Adolf, D., Luetzkendorf, R., Luchtmann, M., & Bernarding, J. (2012). Building virtual reality fMRI paradigms: A framework for presenting immersive virtual environments. *Journal of*

Neuroscience Methods, *209*(2), 290–298. https://doi.org/10.1016/j.jneumeth.2012.06.025

Muhle-Karbe, P. S., Andres, M. & Brass, M. (2014). Transcranial magnetic stimulation dissociates prefrontal and parietal contributions to task preparation. *Journal of Neuroscience*, *34*(37), 12481–12489. https://doi.org/10.1523/JNEUROSCI.4931-13.2014

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, *25*(3), 653–660. https://doi.org/10.1016/j.neuroimage.2004.12.005

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*. https://doi.org/10.1038/s41593-019-0470-8

Norman, D. A., & Shallice, T. (1980). Attention to action: willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D, Shapiro (Eds.), Consciousness and self-regulation: advances in research and theory. New York: Plenum Press.

Oliver, D., Tachtsidis, I., & Hamilton, A. de C. (2017). The role of parietal cortex in overimitation: a study with fNIRS. *Social Neuroscience*, *00*(00), 1–12. https://doi.org/10.1080/17470919.2017.1285812

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Palenciano, A. F., Díaz-Gutiérrez, P., González-García, C., & Ruz, M. (2017).

Neural mechanisms of cognitive control / Mecanismos neurales de control cognitivo. *Estudios de Psicología*, *38*(2), 311–337. https://doi.org/10.1080/02109395.2017.1305060

Palenciano, A. F., González-García, C., Arco, J. E., Pessoa, L., & Ruz, M. (2019a). Representational organization of novel task sets during proactive encoding. *The Journal of Neuroscience*, *39*(42), 8386-8397. https://doi.org/10.1523/jneurosci.0725-19.2019

Palenciano, A. F., González-García, C., Arco, J. E., & Ruz, M. (2019b). Transient and Sustained Control Mechanisms Supporting Novel Instructed Behavior. *Cerebral Cortex*, *29*(9), 3948–3960. https://doi.org/10.1093/cercor/bhy273

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classi!ers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007.Machine

Petersen, S. E., & Dubis, J. W. (2012). The mixed block/event-related design. *NeuroImage*, *62*(2), 1177–1184. https://doi.org/10.1016/j.neuroimage.2011.09.084

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1–25. https://doi.org/10.1111/nyas.13948

Pleger, B., & Timmann, D. (2018). The role of the human cerebellum in linguistic prediction, word generation and verbal working memory: evidence from brain imaging, non-invasive cerebellar stimulation and lesion studies. *Neuropsychologia, 115,* 204-210. https://doi.org/10.1016/j.neuropsychologia.2018.03.012

Posner, M., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42. https://doi.org/10.1007/s11069-007-9150-1

Posner, M. I., & Snyder, C. R. R. (1975). Attention and Cognitive Control. In R. L. Solso (Ed.), *Information Processing and Cognition* (pp. 55-85). Hillsdale, NJ: Erlbaum.

Puri, A. M., Wojciulik, E., & Ranganath, C. (2009). Category expectation modulates baseline and stimulus-evoked activity in human inferotemporal cortex. *Brain Research*, *1301*, 89–99. https://doi.org/10.1016/j.brainres.2009.08.085

Qiao, L., Zhang, L., Chen, A., & Egner, T. (2017). Dynamic Trial-by-Trial Recoding of Task-Set Representations in the Frontoparietal Cortex Mediates Behavioral Flexibility. *The Journal of Neuroscience*, *37*(45), 11037–11050. https://doi.org/10.1523/jneurosci.0935-17.2017

Raichle, M. (2015). The Brain's Default Network. *Annals of the New York Academy of Sciences*, *38*, 433–447. https://doi.org/10.1196/annals.1440.011

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of anew (Affective Norms for English Words). *Behavior Research Methods*, *39*(3), 600–605. https://doi.org/10.3758/BF03193031

Reggente, N., Essoe, J. K. Y., Aghajan, Z. M., Tavakoli, A. V., McGuire, J. F., Suthana, N. A., & Rissman, J. (2018). Enhancing the ecological validity of fMRI memory research using virtual reality. *Frontiers in Neuroscience*, *12*, 408. https://doi.org/10.3389/fnins.2018.00408

Reverberi, C., Görgen, K., & Haynes, J. D. (2012). Compositionality of Rule Representations in Human Prefrontal Cortex. *Cerebral Cortex*, *22*(6),

1237–1246. https://doi.org/10.1093/cercor/bhr200

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447. https://doi.org/10.1126/science.1100301

Rikhye, R. V., Gilra, A., & Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, *21*(12), 1753–1763. https://doi.org/10.1038/s41593-018-0269-z

Rilling, J. K., & Sanfey, A. G. (2011). The Neuroscience of Social Decision-Making. *Annual Review of Psychology*, *62*(1), 23–48. https://doi.org/10.1146/annurev.psych.121208.131647

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/http://dx.doi.org/10.1037/0096-3445.124.2.207

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 763–797. https://doi.org/10.1037/0096-1523.27.4.763

Rushworth, M. F. S., Buckley, M. J., Gough, P. M., Alexander, I. H., Kyriazis, D., McDonald, K. R., & Passingham, R. E. (2005). Attentional selection and action selection in the ventral and orbital prefrontal cortex. *Journal of Neuroscience*, *25*(50), 11628–11636. https://doi.org/10.1523/JNEUROSCI.2765-05.2005

Ruz, M., Madrid, E., & Tudela, P. (2013). Interactions between perceived emotions and executive attention in an interpersonal game. *Social Cognitive and Affective Neuroscience*, *8*(7), 838–844.

https://doi.org/10.1093/scan/nss080

Ruz, M., Moser, A., & Webster, K. (2011). Social expectations bias decision-making in uncertain inter-personal situations. *PLoS ONE*, *6(2): e157*. https://doi.org/10.1371/journal.pone.0015762

Ruz, M., & Tudela, P. (2011). Emotional conflict in interpersonal interactions. *NeuroImage*, *54*(2), 1685–1691. https://doi.org/10.1016/j.neuroimage.2010.08.039

Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, *31*(1), 219–245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature Neuroscience*, *6*(1), 75–81. https://doi.org/10.1038/nn987

Sakai, K., & Passingham, R. E. (2006). Prefrontal Set Activity Predicts Rule-Specific Neural Processing during Subsequent Cognitive Performance. *Journal of Neuroscience*, *26*(4), 1211–1218. https://doi.org/10.1523/JNEUROSCI.3887-05.2006

Sala, J. B., Rämä, P., & Courtney, S. M. (2003). Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory. *Neuropsychologia*, *41*(3), 341–356. https://doi.org/10.1016/S0028-3932(02)00166-5

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science*, (October), 598–602.

Sanfey, A. G. (2009). Expectations and social decision-making: Biasing effects of prior knowledge on Ultimatum responses. *Mind and Society*, *8*(1), 93–107. https://doi.org/10.1007/s11299-009-0053-6

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–1758. https://doi.org/10.1126/science.1082976

Satpute, A. B., & Lindquist, K. A. (2019). The Default Mode Network's Role in Discrete Emotion. *Trends in Cognitive Sciences*, *23*(10), 851–864. https://doi.org/10.1016/j.tics.2019.07.003

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235–239. https://doi.org/10.1016/j.conb.2006.03.001

Saxe, R, & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime user's guide. Pittsburgh: Psychology Software Tools Inc.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE*, *4*(3). https://doi.org/10.1371/journal.pone.0004869

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Schwarz, K. A., Pfister, R., & Büchel, C. (2016). Rethinking Explicit Expectations: Connecting Placebos, Social Cognition, and Contextual Perception. *Trends in Cognitive Sciences*, *20*(6), 469–480. https://doi.org/10.1016/j.tics.2016.04.001

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H.,

Reiss, A.L. & Greicius, M. D. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. *Journal of Neuroscience*, *27*(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. https://doi.org/10.1037/h0027448

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain Responses to the Acquired Moral Status of Faces. *Neuron*, *41*(4), 653–662. https://doi.org/10.1016/S0896-6273(04)00014-5

Smith, R. E. (2003). The Cost of Remembering to Remember in Event-Based Prospective Memory: Investigating the Capacity Demands of Delayed Intention Performance. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(3), 347–361. https://doi.org/10.1037/0278-7393.29.3.347

Smith, V., Mitchell, D. J., & Duncan, J. (2018). Role of the Default Mode Network in Cognitive Transitions, *Cerebral Cortex, 28*(10), 3685–3696. https://doi.org/10.1093/cercor/bhy167

Sokolov, A. A., Miall, R. C., & Ivry, R. B. (2017). The Cerebellum: Adaptive Prediction for Movement and Cognition. *Trends in Cognitive Sciences*, *21*(5), 313–332. https://doi.org/10.1016/j.tics.2017.02.005

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545. https://doi.org/10.1038/nn.2112

Spreng, N. R. (2012). The fallacy of a "task-negative" network. *Frontiers in Psychology, 3*, 145. https://doi.org/10.3389/fpsyg.2012.00145

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2008). The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510. https://doi.org/10.1162/jocn.2008.21029

Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, *105*(34), 12569–12574. https://doi.org/10.1073/pnas.0800005105

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(6), 795–797. https://doi.org/10.1038/nn.4296

Stolier, R. M., & Freeman, J. B. (2017). A Neural Mechanism of Social Categorization. *The Journal of Neuroscience*, *37*(23), 5711–5721. https://doi.org/10.1523/JNEUROSCI.3334-16.2017

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Sulpizio, V., Committeri, G., Lambrey, S., Berthoz, A., & Galati, G. (2013). Selective role of lingual/parahippocampal gyrus and retrosplenial complex

in spatial memory across viewpoint changes relative to the environmental reference frame. *Behavioural Brain Research*, *242*(1), 62–75. https://doi.org/10.1016/j.bbr.2012.12.031

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive Codes for Forthcoming Perception in the Frontal Cortex. *Science*, *314*(5803), 1311–1314. https://doi.org/10.1126/science.1132028

Szameitat, A. J., Schubert, T., Müller, K., & Von Yves Cramon, D. (2002). Localization of executive functions in dual-task performance with fMRI. *Journal of Cognitive Neuroscience*, *14*(8), 1184–1199. https://doi.org/10.1162/089892902760807195

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(1), 194–199. https://doi.org/10.1073/pnas.1511905112

Telga, M., de Lemus, S., Cañadas, E., Rodríguez-Bailón, R., & Lupiáñez, J. (2018). Category-based learning about deviant outgroup members hinders performance in trust decision making. *Frontiers in Psychology*, *9*, 1008. https://doi.org/10.3389/fpsyg.2018.01008

Thornton, M. A., & Mitchell, J. P. (2017). Theories of Person Perception Predict

Patterns of Neural Activity During Mentalizing. *Cerebral Cortex*, 1–16. https://doi.org/10.1093/cercor/bhx216

Thye, M. D., Murdaugh, D. L., & Kana, R. K. (2018). Brain Mechanisms Underlying Reading the Mind from Eyes, Voice, and Actions. *Neuroscience*, *374*, 172–186. https://doi.org/10.1016/j.neuroscience.2018.01.045

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology*, *23*(3), 373–380. https://doi.org/10.1016/j.conb.2012.12.010

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509. https://doi.org/10.1146/annurev-psych-120710-100412

Torres-Quesada, M., Korb, F. M., Funes, M. J., Lupiáñez, J., & Egner, T. (2014). Comparing neural substrates of emotional vs. non-emotional conflict modulation by global control context. *Frontiers in Human Neuroscience*, *8*, 66. https://doi.org/10.3389/fnhum.2014.00066

Tortosa, M. I., Lupiáñez, J., & Ruz, M. (2013). Race, emotion and trust: An ERP study. *Brain Research*, *1494*, 44–55. https://doi.org/10.1016/j.biopsycho.2008.03.004

Tortosa, M. I., Strizhko, T., Capizzi, M., & Ruz, M. (2013). Interpersonal effects of emotion in a multi-round Trust Game. *Psicologica*, *34*(2), 179–198.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–

249. https://doi.org/http://doi.org/10.1016/j.psychres.2008.05.006

Tschentscher, N., Mitchell, D., & Duncan, J. (2017). Fluid Intelligence Predicts Novel Rule Implementation in a Distributed Frontoparietal Control Network. *The Journal of Neuroscience*, *37*(18), 4841–4847. https://doi.org/10.1523/jneurosci.2478-16.2017

Tsetsos, K., Wyart, V., Shorkey, S. P., & Summerfield, C. (2014). Neural mechanisms of economic commitment in the human medial prefrontal cortex. *ELife*, *3*, 1–17. https://doi.org/10.7554/eLife.03701

Turner, B. (2010). Comparison of methods for the use of pattern classificaion on rapid event-related fMRI data. Poster session presented at the Annual Meeting of the Society for Neuroscience, San Diego, CA.

van der Wel, P. & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review, 25*, 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (1st ed., Vol. 42, pp. 45–96). Burlington, MA: Academic Press. https://doi.org/10.1016/S0065-2601(10)42002-X

van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Van Overwalle, F., Baetens, K., Mariën, P., & Vandekerckhove, M. (2014). Social cognition and the cerebellum: A meta-analysis of over 350 fMRI studies. *NeuroImage*, *86*, 554–572. https://doi.org/10.1016/j.neuroimage.2013.09.033

Visscher, K. M., Miezin, F. M., Kelly, J. E., Buckner, R. L., Donaldson, D. I., McAvoy, M. P., Bhalodia, V.M. & Petersen, S. E. (2003). Mixed blocked/event-related designs separate transient and sustained activity in fMRI. *NeuroImage*, *19*(4), 1694–1708. https://doi.org/10.1016/S1053-8119(03)00178-2

Vogt, B. A. (2016). Midcingulate cortex: Structure, connections, homologies, functions and diseases. *Journal of Chemical Neuroanatomy*, *74*, 28–46. https://doi.org/10.1016/j.jchemneu.2016.01.010

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron*, *31*(2), 329–338. https://doi.org/10.1016/S0896-6273(01)00359-2

Wang, W.-H., Shih, Y.-H., Yu, H.-Y., Yen, D.-J., Lin, Y.-Y., Kwan, S.-Y., … Hua, M.-S. (2015). Theory of mind and social functioning in patients with temporal lobe epilepsy. *Epilepsia*, *56*(7), 1117–1123. https://doi.org/10.1111/epi.13023

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal Representations of Task Context Support the Flexible Control of Goal-Directed Cognition. *Journal of Neuroscience*, *34*(32), 10743–10755. https://doi.org/10.1523/JNEUROSCI.5282-13.2014

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011a). Adaptive Coding of Task-Relevant Information in Human Frontoparietal Cortex. *Journal of Neuroscience*, *31*(41), 14592–14599.

https://doi.org/10.1523/JNEUROSCI.2616-11.2011

Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. *Journal of Cognitive Neuroscience*, *28*(10), 1433–1454. https://doi.org/10.1162/jocn_a_00981

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011b). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752. https://doi.org/10.1016/j.neuroimage.2010.04.035

Yeung, N., & Sanfey, A. G. (2004). Independent Coding of Reward Magnitude and Valence in the Human Brain. *Journal of Neuroscience*, *24*(28), 6258–6264. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Zhen, Z., Fang, H., & Liu, J. (2013). The Hierarchical Brain Network for Face Recognition. *PLoS ONE*, *8*(3): e59886. https://doi.org/10.1371/journal.pone.0059886